CHAPTER 2: Physics-Based Derivation of the I-V Model

BSIM3v3's development is based upon finding solutions to Poisson's equation using Gradual Channel Approximation (GCA) and Quasi-Two Dimensional Approximation (QTDA) approaches. It includes compact, analytical expressions for the following physical phenomenon observed in present day MOS devices [1]:

- Short and narrow channel effects on threshold voltage.
- Non-uniform doping effect (in both lateral and vertical directions).
- Mobility reduction due to vertical field.
- Bulk charge effect.
- Carrier velocity saturation.
- Drain-induced barrier lowering (*DIBL*).
- Channel length modulation (*CLM*).
- Substrate current induced body effect (*SCBE*).
- Subthreshold conduction.
- Source/drain parasitic resistances.

2.1 Non-Uniform Doping and Small Channel Effects on Threshold Voltage

Accurate modeling of threshold voltage (Vth) is one of the most important requirements for the precise description of a device's electrical characteristics. In addition, it serves as a useful reference point for the evaluation of device operation

regimes. By using threshold voltage, the whole device operation regime can be divided into three operational regions.

First, if the gate voltage is greater than the threshold voltage, the inversion charge density is larger than the substrate doping concentration. The MOSFET is then operating in the strong inversion region and drift current is dominant. Second, if the gate voltage is much less than the threshold voltage, the inversion charge density is smaller than the substrate doping concentration. The MOSFET is now considered to be operating in the weak inversion (or subthreshold) region. Diffusion current is now dominant [2]. Lastly, if the gate voltage is very close to the threshold voltage, the inversion charge density is close to the doping concentration and the MOSFET is operating in the transition region. In such a case, both diffusion and drift currents are equally important.

The standard threshold voltage of a MOSFET with long channel length/width and uniform substrate doping concentration [2] is given by:

$$V_{th} = V_{FB} + \phi_s + \gamma\sqrt{\phi_s - V_{bs}} = V_{Tideal} + \gamma(\sqrt{\phi_s - V_{bs}} - \sqrt{\phi_s}) \tag{2.1.1}$$

where V_{FB} is the flat band voltage, V_{Tideal} is the ideal threshold voltage of the long channel device at zero volt substrate bias, and γ is the substrate bias effect coefficient and is given by:

$$\gamma = \frac{\sqrt{2\varepsilon_{si} q N_a}}{C_{ox}} \tag{2.1.2}$$

where N_a is the substrate doping concentration. The surface potential is given by:

$$\phi_s = 2\frac{K_B T}{q}\ln(\frac{N_a}{n_i}) \tag{2.1.3}$$

Equation (2.1.1) assumes that the channel is uniform and makes use of the one dimensional Poisson equation in the vertical direction of the channel. This model is valid only when the substrate doping concentration is constant and the channel length is long. Under these conditions, the potential is uniform along the channel. But in reality, these two conditions are not always satisfied. Modifications have to be made when the substrate doping concentration is not uniform or and when the channel length is short, narrow, or both.

2.1.1 Vertical Non-Uniform Doping Effect

The substrate doping level is not constant in the vertical direction as shown in Figure 2-1.

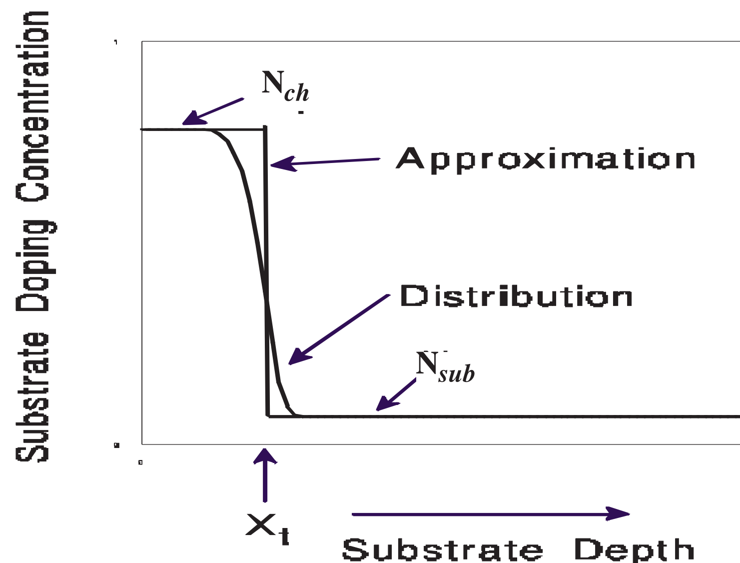

Figure 2-1. Actual substrate doping distribution and its approximation.

The substrate doping concentration is usually higher near the silicon to silicon dioxide interface (due to the threshold voltage adjust implant) than deep into substrate. The distribution of impurity atoms inside the substrate is approximately a half gaussian distribution, as shown in Figure 2-1. This non-uniformity will make γ in Eq. (2.1.2) a function of the substrate bias. If the depletion width is less than X_t as shown in Figure 2-1, N_a in Eq. (2.1.2) is equal to N_{ch}, otherwise it is equal to N_{sub}.

In order to take into account such non-uniform substrate doping, the following threshold voltage model is proposed:

$$V_{th} = V_{Tideal} + K_1(\sqrt{\phi_s - V_{bs}} - \sqrt{\phi_s}) - K_2 V_{bs} \tag{2.1.4}$$

For a zero substrate bias, Eqs. (2.1.1) and (2.1.4) give the same result. K_1 and K_2 can be determined by the criteria that V_{th} and its derivative versus V_{bs} should be the same at V_{bm}, where V_{bm} is the maximum substrate bias voltage. Therefore, using equations (2.1.1) and (2.1.4), K_1 and K_2 [3] will be given by the following:

$$K_1 = \gamma_2 - 2K_2\sqrt{\phi_s - V_{bm}} \tag{2.1.5}$$

$$K_2 = (\gamma_1 - \gamma_2)\frac{\sqrt{\phi_s - V_{bx}} - \sqrt{\phi_s}}{2\sqrt{\phi_s}(\sqrt{\phi_s - V_{bm}} - \sqrt{\phi_s}) + V_{bm}} \tag{2.1.6}$$

where γ_1 (or γ_2) is the body effect coefficient when the substrate doping concentration is N_{ch} (or N_{sub}) as shown in Figure 2-1.

 BSIM3v3 Manual Copyright © 1995, 1996, UC Berkeley

$$\gamma_1 = \frac{\sqrt{2q\varepsilon_{si} N_{ch}}}{C_{ox}} \tag{2.1.7}$$

$$\gamma_2 = \frac{\sqrt{2q\varepsilon_{si} N_{sub}}}{C_{ox}} \tag{2.1.8}$$

V_{bx} is the body bias when the depletion width is equal to X_t. Therefore, V_{bx} satisfies:

$$\frac{qN_{ch} X_t^2}{2\varepsilon_{si}} = \phi_s - V_{bx} \tag{2.1.9}$$

If the devices are available, K_1 and K_2 can be determined experimentally. If the devices are not available but the user knows the doping concentration distribution, the user can input the appropriate parameters to specify doping concentration distribution (e.g. N_{ch}, N_{sub}, X_t). Then, K_1 and K_2 can be calculated using equations (2.1.5) and (2.1.6).

2.1.2 Lateral Non-Uniform Doping Effect:

For some technologies, the doping concentration near the drain and the source is higher than that in the middle of the channel. This is referred to as lateral non-uniform doping and is shown in Figure 2-2. As the channel length becomes shorter, lateral non-uniform doping will cause the threshold voltage to increase in magnitude because the average doping concentration in the channel is larger. The average channel doping can be calculated as follows:

$$N_{eff} = \frac{N_a(L-2L_x)+N_{ds}\cdot 2L_x}{L} = N_a(1+\frac{2L_x}{L}\frac{N_{ds}-N_a}{N_a}) \equiv N_a(1+\frac{Nlx}{L}) \tag{2.1.10}$$

Due to this lateral non-uniform doping effect, Eq. (2.1.4) becomes:

$$V_{th} = V_{tho} + K_1(\sqrt{\Phi_s - V_{bs}} - \sqrt{\Phi_s}) - K_2 V_{bs} + K_1\left(\sqrt{1+\frac{N_{LX}}{L_{eff}}} - 1\right)\sqrt{\Phi_s} \tag{2.1.11}$$

Eq. (2.1.11) can be derived by setting $V_{bs} = 0$, and using $K_1 \propto (N_{eff})^{0.5}$. The fourth term in Eq. (2.1.11) is used empirically to model the body bias dependence of the lateral non-uniform doping effect. This effect gets stronger at a lower body bias. Examination of Eq. (2.1.11) shows that the threshold voltage will increase as channel length decreases [3].

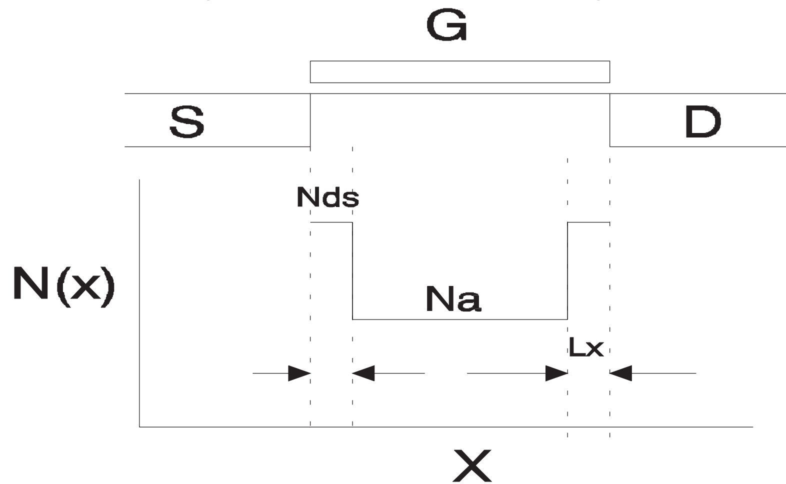

Figure 2-2. Doping concentration along the channel is non-uniform.

 BSIM3v3 Manual Copyright © 1995, 1996, UC Berkeley

2.1.3 Short Channel Effect

The threshold voltage of a long channel device is independent of the channel length and the drain voltage. Its dependence on the body bias is given by Eq. (2.1.4). However, as the channel length becomes shorter and shorter, the threshold voltage shows a greater dependence on the channel length and the drain voltage. The dependence of the threshold voltage on the body bias becomes weaker as channel length becomes shorter, because the body bias has less control of the depletion region. Short-channel effects must be included in the threshold voltage in order to model deep-submicron devices correctly.

The short channel effect can be modeled in the threshold voltage by the following:

$$V_{th} = V_{tho} + K_1(\sqrt{\Phi_s - V_{bs}} - \sqrt{\Phi_s}) - K_2 V_{bs} + K_1\left(\sqrt{1 + \frac{N_{LX}}{L_{eff}}} - 1\right)\sqrt{\Phi_s} - \Delta V_{th} \quad (2.1.12)$$

where ΔV_{th} is the threshold voltage reduction due to the short channel effect. Many models have been developed to calculate ΔV_{th}. They used either numerical solutions [4], a two-dimensional charge sharing approach [5,6], or a simplified Poisson's equation in the depletion region [7-9]. A simple, accurate, and physical model was developed by Z.H. Liu, et al, [10]. Their model was derived by solving the quasi two-dimension Poisson equation along the channel. This quasi-2D model concluded that:

$$\Delta V_{th} = \theta_{th}(L)[2(V_{bi} - \phi_s) + V_{ds}] \quad (2.1.13)$$

where V_{bi} is the built-in voltage of a PN junction between the substrate and the source. V_{bi} is given by:

$$V_{bi} = \frac{K_B T}{q} \ln\left(\frac{N_{ch} N_d}{n_i^2}\right) \tag{2.1.14}$$

where N_d in Eq. (2.1.14) is the source doping concentration, and N_{ch} is the substrate doping concentration. The expression $\theta_{th}(L)$ is a short channel effect coefficient that has a strong dependence on the channel length and is given by:

$$\theta_{th}(L) = [\exp(-L/2l_t) + 2\exp(-L/l_t)] \tag{2.1.15}$$

l_t is referred to as the "characteristic length" and is given by:

$$l_t = \sqrt{\frac{\varepsilon_{si} T_{ox} X_{dep}}{\varepsilon_{ox} \eta}} \tag{2.1.16}$$

X_{dep} is the depletion width in the substrate and is given by:

$$X_{dep} = \sqrt{\frac{2\varepsilon_{si}(\phi_s - V_{bs})}{qN_{ch}}} \tag{2.1.17}$$

X_{dep} is larger near the drain than in the middle of the channel due to the drain voltage. X_{dep} / η represents the average depletion width along the channel.

Based on above discussion, the influences of drain/source charge sharing and DIBL effects to Vth are described by (2.1.15). However, in order to

 BSIM3v3 Manual Copyright © 1995, 1996, UC Berkeley

make the model fit a wide technology range, several parameters such as D_{vt0}, D_{vt2}, *Dsub*, *Eta0* and *Etab* are introduced, and the following modes are used in SPICE to account for charge sharing and the DIBL effects separately.

$$\theta_{th(L)} = D_{vt0}[\exp(-D_{vt1} L / 2l_t) + 2\exp(-D_{vt1} L / l_t)] \tag{2.1.18}$$

$$\Delta V_{th(L)} = \theta_{th(L)}(V_{bi} - \phi_s) \tag{2.1.19}$$

$$l_t = \sqrt{\frac{\varepsilon_{si} T_{ox} X_{dep}}{\varepsilon_{ox}}}(1 + D_{vt2} V_{bs}) \tag{2.1.20}$$

$$\theta_{dibl(L)} = [\exp(-D_{sub} L / 2l_{t0}) + 2\exp(-D_{sub} L / l_{t0})] \tag{2.1.21}$$

$$\Delta V_{th}(V_{ds}) = \theta_{dibl(L)}(E_{ta0} + E_{tab} V_{bs}) V_{ds} \tag{2.1.22}$$

where l_{t0} is calculated by Eq. (2.1.20) at zero body-bias. D_{vt1} is basically equal to $1/(\eta)^{1/2}$ in Eq. (2.1.16). D_{vt2} is introduced to take care of the dependence of the doping concentration on substrate bias since the doping concentration is not uniform in the vertical direction of the channel. X_{dep} is calculated using the doping concentration in the channel (N_{ch}). D_{vt0}, D_{vt1},D_{vt2}, *Eta0*, *Etab* and *Dsub*, which are determined experimentally, can improve accuracy greatly. Even though Eqs. (2.1.18), (2.1.21) and (2.1.15) have different coefficients, they all still have the same functional forms. This the device physics represented by Eqs. (2.1.18), (2.1.21) and (2.1.15) are still the same.

As channel length L decreases, ΔV_{th} will increase, and in turn V_{th} will decrease. If a MOSFET has a *LDD* structure, N_d in Eq. (2.1.14) is the doping concentration in the lightly doped region. V_{bi} in a *LDD*-MOSFET will be smaller as compared to conventional MOSFETs, therefore the threshold voltage reduction due to the short channel effect will be smaller in *LDD*-MOSFETs.

As the body bias becomes more negative, the depletion width will increase as shown in Eq. (2.1.17). Hence ΔV_{th} will increase due to the increase in l_t. The term:

$$V_{Tideal} + K_1\sqrt{\phi_s - V_{bs}} - K_2 V_{bs}$$

will also increase as V_{bs} becomes more negative (for NMOS). Therefore, the changes in:

$$V_{Tideal} + K_1\sqrt{\phi_s - V_{bs}} - K_2 V_{bs}$$

and in ΔV_{th} will compensate for each other and make V_{th} less sensitive to V_{bs}. This compensation is more significant as the channel length is shortened. Hence, the V_{th} of short channel MOSFET is less sensitive to body bias as compared to a long channel MOSFET. For the same reason, the *DIBL* effect and the channel length dependence of V_{th} are stronger as V_{bs} is made more negative. This was verified by experimental data shown in Figure 2-3 and Figure 2-4. Although Liu, et al, found a accelerated V_{th} roll-off and non-linear drain voltage dependence [10] as the channel became very short, a linear dependence of V_{th} on V_{ds} is nevertheless a good approximation for circuit simulation as shown in Figure 2-4. This figure shows that Eq. (2.1.13) can fit the experimental data very well.

 BSIM3v3 Manual Copyright © 1995, 1996, UC Berkeley

Furthermore, Figure 2-5 shows how this model for V_{th} can fit various channel lengths under various bias conditions.

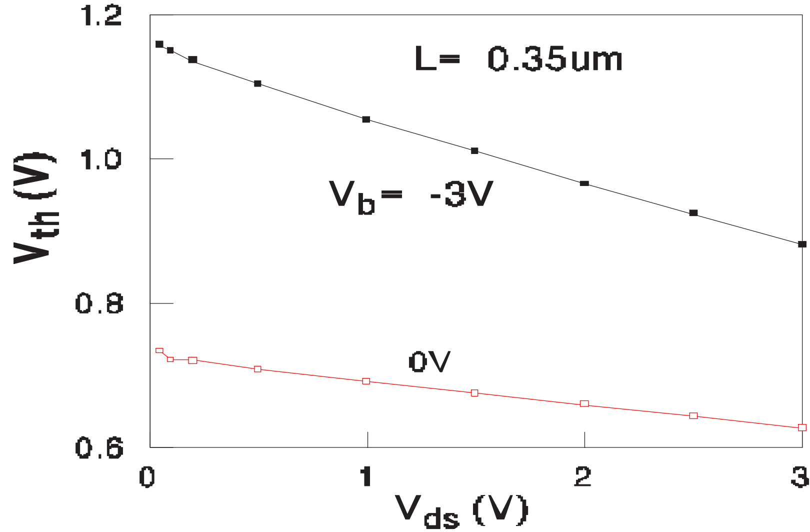

Figure 2-3. Threshold voltage versus the drain voltage at different body biases.

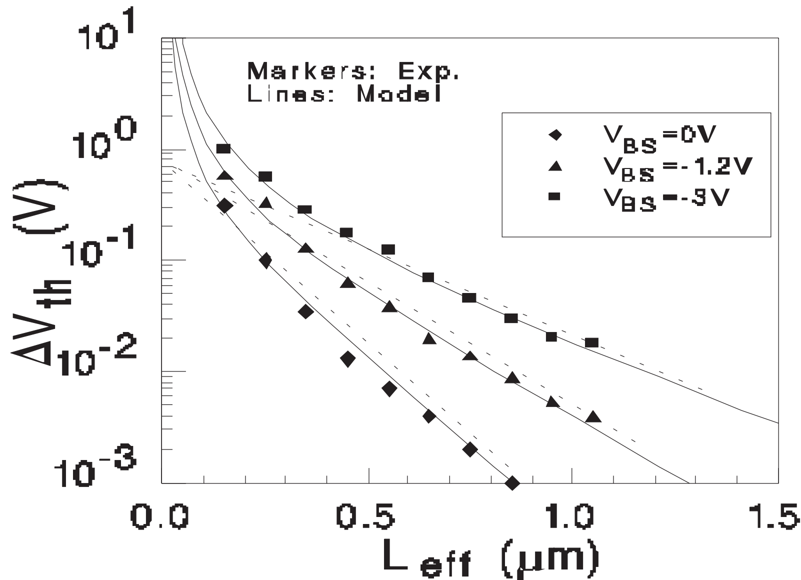

Figure 2-4. Channel length dependence of threshold voltage.

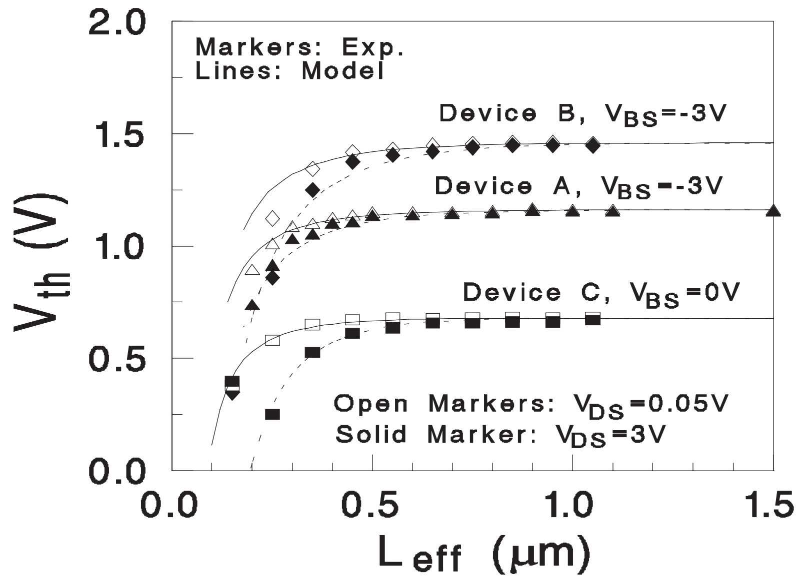

Figure 2-5. Threshold voltage versus channel length at different biases.

2.1.4 Narrow Channel Effect

The actual depletion region in the channel is always larger than what is usually assumed under the one-dimensional analysis due to the existence of fringing fields [2]. This effect becomes very substantial as the channel width decreases and the depletion region underneath the fringing field becomes comparable to the "classical" depletion layer formed from the vertical field. The net result is an increase threshold voltage magnitude. It is shown in [2] that this increase can be modeled as:

$$\frac{\pi q N_a (X_{d\max})^2}{2C_{ox}W} = 3\pi \frac{T_{ox}}{W}\phi_s \tag{2.1.23}$$

 BSIM3v3 Manual Copyright © 1995, 1996, UC Berkeley

The right hand side of Eq. (2.1.23) represents the additional voltage increase. BSIM3v3 models this change in threshold voltage by Eq. (2.1.24a). This formulation includes but is not limited to the inverse of channel width due to the fact that the overall narrow width effect is dependent on process (i.e. isolation technology) as well. Hence, the introduction of parameters K_3, K_{3b}, and W_0.

$$(K_3 + K_{3b} V_{bs}) \frac{T_{ox}}{(W_{eff}' + W_0)} \phi_s \qquad (2.1.24a)$$

W_{eff}' is the effective channel width (with no bias dependencies), which will be defined Section 2.9. In addition, we must also consider the narrow width effect for small channel lengths. To do this we introduce the following:

$$D_{VT0w} \left(exp(-D_{VT1w} \frac{W_{eff}' L_{eff}}{2 l_{tw}}) + 2 exp(-D_{VT1w} \frac{W_{eff}' L_{eff}}{l_{tw}}) \right) (V_{bi} - \Phi_s) \qquad (2.1.24b)$$

When all of the above considerations for non-uniform doping, short and narrow channel effects on threshold voltage are considered, the final, complete Vth expression implemented in SPICE is as follows:

$$\begin{aligned} V_{th} = {} & V_{tho} + K_1 (\sqrt{\Phi_s - V_{bseff}} - \sqrt{\Phi_s}) - K_2 V_{bseff} \\ & + K_1 \left(\sqrt{1 + \frac{N_{LX}}{L_{eff}}} - 1 \right) \sqrt{\Phi_s} + (K_3 + K_{3b} V_{bseff}) \frac{T_{OX}}{W_{eff}' + W_0} \Phi_s \\ & - D_{VT0w} \left(exp(-D_{VT1w} \frac{W_{eff}' L_{eff}}{2 l_{tw}}) + 2 exp(-D_{VT1w} \frac{W_{eff}' L_{eff}}{l_{tw}}) \right) (V_{bi} - \Phi_s) \\ & - D_{VT0} \left(exp(-D_{VT1} \frac{L_{eff}}{2 l_t}) + 2 exp(-D_{VT1} \frac{L_{eff}}{l_t}) \right) (V_{bi} - \Phi_s) \end{aligned} \qquad (2.1.25)$$

(2.1.25) (cont.)

$$- \left(\exp(-D_{sub}\frac{L_{eff}}{2l_{to}}) + 2\exp(-D_{sub}\frac{L_{eff}}{l_{to}}) \right) (E_{tao} + E_{tab} V_{bseff})\, V_{ds}$$

In Eq. (2.1.25), all Vbs terms have been substituted with a Vbseff expression as shown in Eq. (2.1.26). This is done in order to set an upper bound for the body bias value during simulations. Unreasonable values can occur if this expression is not introduced. See Section 3.8 for details

(2.1.26).

$$V_{bseff} = V_{bc} + 0.5[V_{bs} - V_{bc} - \delta_1 + \sqrt{(V_{bs} - V_{bc} - \delta_1)^2 - 4\delta_1 V_{bc}}\ ]$$

where δ_1=0.001. The parameter *Vbc* is the maximum allowable V_{bs} value and is calculated from the condition of $dV_{th}/dV_{bs}=0$ for the V_{th} expression of 2.1.4, 2.1.5, and 2.1.6 is equal to:

$$V_{bc} = 0.9(\phi s - \frac{K1^2}{4K2^2})$$

2.2 Mobility Model

A good model for surface carrier mobility is very critical to the accuracy of a MOSFET model. The scattering mechanisms responsible for surface mobility basically include phonons, coulombic scattering sites, and surface roughness [11, 12]. For good quality interfaces, phonon scattering is generally the dominant scattering mechanism at room temperature. In general, mobility depends on many process parameters and bias conditions. For example, mobility depends on gate oxide thickness, doping concentration, threshold voltage, gate voltage and

 BSIM3v3 Manual Copyright © 1995, 1996, UC Berkeley

substrate voltage, etc. Sabnis and Clemens [13] proposed an empirical unified formulation based on the concept of an effective field E_{eff} which lumps many process parameters and bias conditions together. E_{eff} is defined by

$$E_{eff} = \frac{Q_B + (Q_n/2)}{\varepsilon_{si}} \tag{2.2.1}$$

The physical meaning of E_{eff} can be interpreted as the average electrical field experienced by the carriers in the inversion layer [14]. The unified formulation of mobility is then empirically given by:

$$\mu_{eff} = \frac{\mu_0}{1 + (E_{eff}/E_0)^{\nu}} \tag{2.2.2}$$

Values for μ_0, E_0, and ν were reported by Liang *et al*. [15] and Toh *et al*. [16] to be the following for electrons and holes:

Parameter	Electron (surface)	Hole (surface)
$\mu_0\ (cm^2/V)$	670	160
$E_0\ (MV/cm)$	0.67	0.7
ν	1.6	1.0

Table 2-1. Mobility and related parameters for electrons and holes.

For a NMOS transistor with n-type poly-silicon gate, Eq. (2.2.1) can be rewritten in a more useful form that explicitly relates E_{eff} to the device parameters [14]:

$$E_{eff} \cong \frac{V_{gs} + V_{th}}{6T_{ox}} \tag{2.2.3}$$

Eq. (2.2.2) fits experimental data very well [15], but it involves a power function which is a very time consuming function for circuit simulators such as SPICE. A Taylor expansion Eq. (2.2.2) is used, and the coefficients are left to be determined using experimental data or to be obtained by fitting the unified formulation. Thus, we have:

(Mobmod=1) (2.2.4)

$$\mu_{eff} = \frac{\mu_o}{1 + (U_a + U_c V_{bseff})(\frac{V_{gst} + 2V_{th}}{T_{OX}}) + U_b(\frac{V_{gst} + 2V_{th}}{T_{OX}})^2}$$

where Vgst=Vgs-Vth. To account for depletion mode devices, another mobility model option is given by the following:

.(Mobmod=2) (2.2.5)

$$\mu_{eff} = \frac{\mu_o}{1 + (U_a + U_c V_{bseff})(\frac{V_{gst}}{T_{OX}}) + U_b(\frac{V_{gst}}{T_{OX}})^2}$$

The unified mobility expressions in subthreshold and strong inversion regions will be discussed in Section 3.2.

To consider the body bias dependence of Eq. 2.2.4 further, we have introduced the following expression:

(For Mobmod=3) (2.2.6)

$$\mu_{eff} = \frac{\mu_o}{1+[U_a(\frac{V_{gst}+2V_{th}}{T_{OX}})+U_b(\frac{V_{gst}+2V_{th}}{T_{OX}})^2](1+U_cV_{bseff})}$$

2.3 Carrier Drift Velocity

Carrier drift velocity is one of the most important parameters that affects device performance characteristics. BSIM3v3 uses a simple and semi-empirical saturation velocity model [17] given by:

(2.3.1)

$$v = \frac{\mu_{eff}E}{1+(E/E_{sat})}, \qquad E < E_{sat}$$

$$= v_{sat}, \qquad E > E_{sat}$$

The parameter E_{sat} corresponds to the critical electrical field at which the carrier velocity becomes saturated. In order to have a continuous velocity model at $E = E_{sat}$, E_{sat} must satisfy:

(2.3.2)

$$E_{sat} = \frac{2v_{sat}}{\mu_{eff}}$$

2.4 Bulk Charge Effect

When the drain voltage is large and/or when the channel length is long, the depletion "thickness" of the channel is not uniform along the channel length. This

will cause threshold voltage to vary along the channel. This effect is called bulk charge effect [14].

In BSIM3v3, the parameter $A_{bulk}$is used to take into account this bulk charge effect. This parameter is a modification from that of BSIM1 and BSIM2 where the bulk charge parameter was "*a*" [3]. Several extracted parameters such as A0, Bo,B1 are introduced to account for the channel length and width dependences of the bulk charge effect. In addition, the parameter *Keta* is introduced to model the change in bilk charge effect under high back or substrate bias conditions. It should be pointed out that narrow width effects have been considered in the formulation of Eq. (2.4.1). The A_{bulk} expression used in BSIM3v3 is given by:

$$A_{bulk} = (1 + \frac{K_1}{2\sqrt{\Phi_s - V_{bseff}}} \{ \frac{A_0 L_{eff}}{L_{eff} + 2\sqrt{X_J X_{dep}}} [1 - A_{gs} V_{gsteff} (\frac{L_{eff}}{L_{eff} + 2\sqrt{X_J X_{dep}}})^2] + \frac{B_o}{Weff' + B_1} \}) \frac{1}{1 + K_{ETA} V_{bseff}} \quad (2.4.1)$$

where A_0, A_{gs}, K_1,B_o, B_1 and Keta are determined by experimental data. Eq. (2.4.1) shows that A_{bulk} is very close to *1.0* if the channel length is small, and A_{bulk} increases as channel length increases.

2.5 Strong Inversion Drain Current (Linear Regime)

2.5.1 Intrinsic Case (Rds=0)

In the strong inversion region, the general current equation at any point y along the channel is given by:

$$I_{ds} = WC_{ox}(V_{gst} - A_{bulk}V_{(y)})v_{(y)} \tag{2.5.1}$$

The parameter $V_{gst} = (V_{gs} - V_{th})$, W is the device channel width, C_{ox} is the gate capacitance per unit area, V(y) is the potential difference between minority-carrier quasi-Fermi potential and the equilibrium Fermi potential in the bulk at point y, $v(y)$ is the velocity of carriers at point y, and A_{bulk} is the coefficient accounting for the bulk charge effect.

With Eq. (2.3.1) (i.e. before carrier velocity saturates), the drain current can be expressed as:

$$I_{ds} = WC_{ox}(V_{gs} - V_{th} - A_{bulk}V_{(y)})\frac{\mu_{eff}E_{(y)}}{1 + E_{(y)}/E_{sat}} \tag{2.5.2}$$

Eq. (2.5.2) can be rewritten as follows:

$$E_{(y)} = \frac{I_{ds}}{\mu_{eff}WC_{ox}(V_{gst} - A_{bulk}V_{(y)}) - I_{ds}/E_{sat}} = \frac{dV_{(y)}}{dy} \tag{2.5.3}$$

By integrating Eq. (2.5.3) from $y = 0$ to $y = L$ and $V(y) = V_s$ to $V(y) = V_d$, we arrive at the following:

$$I_{ds} = \mu_{eff}C_{ox}\frac{W}{L}\frac{1}{1 + V_{ds}/E_{sat}L}(V_{gs} - V_{th} - A_{bulk}V_{ds}/2)V_{ds} \tag{2.5.4}$$

The drain current model in Eq. (2.5.4) is valid before the carrier velocity saturates.

For instances when the drain voltage is high (and thus the lateral electrical field is high at the drain side), the carrier velocity near the drain saturates. The channel now can be reasonably divided into two portions: one adjacent to the source where the carrier velocity is field-dependent and the second where the velocity saturates. At the boundary between these two portions, the channel voltage is the saturation voltage (V_{dsat}) and the lateral electrical is equal to E_{sat}. After the onset of saturation, we can substitute $v = v_{sat}$ and $V_{ds} = V_{dsat}$ into Eq. (2.5.1) to get the saturation current:

$$I_{ds} = WC_{ox}(V_{gst} - A_{bulk}V_{dsat})v_{sat} \tag{2.5.5}$$

By equating eqs. (2.5.4) and (2.5.5) at $E = E_{sat}$ and $V_{ds} = V_{dsat}$, we can solve for saturation voltage V_{dsat}:

$$V_{dsat} = \frac{E_{sat}L(V_{gs} - V_{th})}{A_{bulk}E_{sat}L + (V_{gs} - V_{th})} \tag{2.5.6}$$

2.5.2 Extrinsic Case (Rds>0)

Parasitic source/drain resistance is an important device parameter which can affect MOSFET performance significantly. As a MOSFET's channel length scale down, the parasitic resistance will not be proportionally scaled. As a result, Rds will have a more significant on device characteristics. Model parasitic resistance in a direct method yields a complicated drain current expression. In order to make simulations more efficient, BSIM3v3 models parasitic resistances using simple expressions. The resulting drain current equation in the linear region can be calculate [3] as follows:

 BSIM3v3 Manual Copyright © 1995, 1996, UC Berkeley

(2.5.9)

$$I_{ds} = \frac{V_{ds}}{R_{tot}} = \frac{V_{ds}}{R_{ch} + R_{ds}}$$
$$= \mu_{eff} C_{ox} \frac{W}{L} \frac{1}{1 + V_{ds}/(E_{sat} L)} \frac{(V_{gst} - A_{bulk} V_{ds}/2) V_{ds}}{1 + R_{ds} \mu_{eff} C_{ox} \frac{W}{L} \frac{(V_{gst} - A_{bulk} V_{ds}/2)}{1 + V_{ds}/(E_{sat} L)}}$$

Due to parasitic resistance, the saturation voltage V_{dsat} will be larger than that predicted by Eq. (2.5.6). Let Eq. (2.5.5) be equal to Eq. (2.5.9). The V_{dsat} with parasitic resistance R_{ds} expression is then:

(2.5.10)

$$V_{dsat} = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

The following are the expression for the variables *a*, *b*, and *c*:

(2.5.11)

$$a = A_{bulk}^2 R_{ds} C_{ox} W v_{sat} + (\frac{1}{\lambda} - 1) A_{bulk}$$
$$b = -(V_{gst}(\frac{2}{\lambda} - 1) + A_{bulk} E_{sat} L + 3 A_{bulk} R_{ds} C_{ox} W v_{sat} V_{gst})$$
$$c = E_{sat} L V_{gst} + 2 R_{ds} C_{ox} W v_{sat} V_{gst}^2$$
$$\lambda = A_1 V_{gst} + A_2$$

The last expression for λ is introduced to account for non-saturation effect of the device. In BSIM3v3, parasitic resistance is modeled as:

(2.5.11)

$$R_{ds} = \frac{R_{dsw}[1 + P_{rwg} V_{gsteff} + P_{rwb}(\sqrt{\phi_s - V_{bseff}} - \sqrt{\phi_s})]}{(10^6 W_{eff}')^{Wr}}$$

The variable R_{dsw} is the resistance per unit width, Wr is a fitting parameter, P_{rwb} is the body effect coefficient, and P_{rwg} is the gate-bais effect effect.

2.6 Strong Inversion Current and Output Resistance (Saturation Regime)

A typical I-V curve and its output resistance are shown in Figure 2-6. Considering only the drain current, the I-V curve can be divided into two parts: the linear region in which the drain current increases quickly with the drain voltage and the saturation region in which the drain current has a very weak dependence on the drain voltage. The first order derivative reveals more detailed information about the physical mechanisms which are involved during device operation. The output resistance (which is the reciprocal of the first order derivative of the I-V curve) curve can be clearly divided into four regions in which have distinct R_{out} *vs.* V_{ds} dependences.

The first region is the triode (or linear) region in which carrier velocity is not saturated. The output resistance is very small because the drain current has a strong dependence on the drain voltage. The other three regions belong to the saturation region. As will be discussed later, there are three physical mechanisms which affect the output resistance in the saturation region: channel length modulation (*CLM*) [4, 14], drain-induced barrier lowering (*DIBL*) [4, 6, 14], and the substrate current induced body effect (*SCBE*) [14, 18, 19]. All three mechanisms affect the output resistance in the saturation range, but each of them dominates in only a single region. It will be shown next that channel length modulation (*CLM*)

 BSIM3v3 Manual Copyright © 1995, 1996, UC Berkeley

dominates in the second region, *DIBL* in the third region, and *SCBE* in the fourth region.

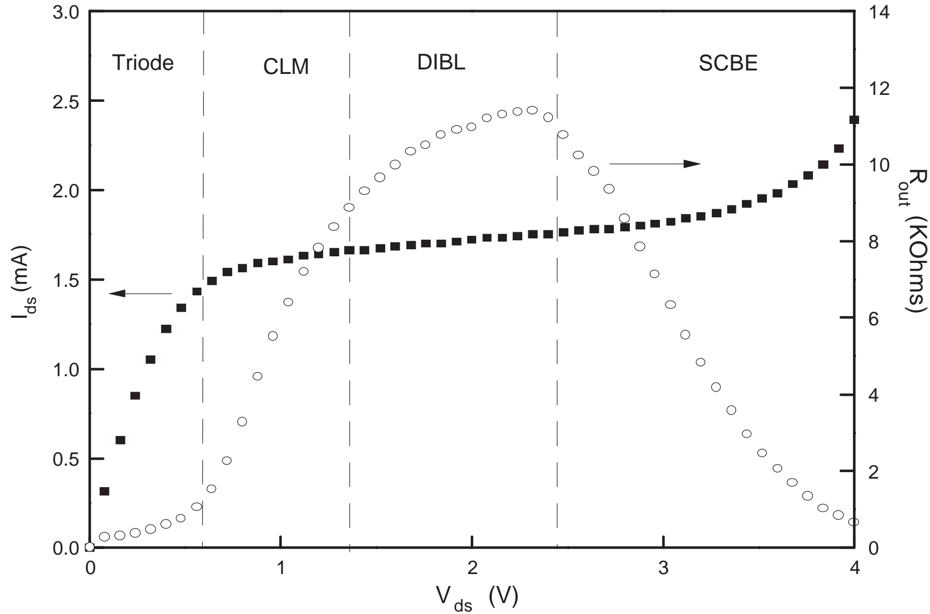

Figure 2-6. General behavior of MOSFET output resistance.

Generally, drain current is a function of the gate voltage and the drain voltage. But the drain current depends on the drain voltage very weakly in the saturation region. A Taylor series can be used to expand the drain current in the saturation region [3].

$$I_{ds}(V_{gs}, V_{ds}) = I_{ds}(V_{gs}, V_{dsat}) + \frac{\partial I_{ds}(V_{gs}, V_{ds})}{\partial V_{ds}}(V_{ds} - V_{dsat})$$
$$\equiv I_{dsat}\left(1 + \frac{V_{ds} - V_{dsat}}{V_A}\right) \quad (2.6.1)$$

where,

(2.6.2)

$$I_{dsat} = I_{ds}(V_{gs}, V_{dsat}) = W v_{sat} C_{ox} (V_{gst} - A_{bulk} V_{dsat})$$

and

(2.6.3)

$$V_A = I_{dsat} \left(\frac{\partial I_{ds}}{\partial V_{ds}} \right)^{-1}$$

The parameter V_A is called the Early Voltage (analogous to the BJT) and is introduced for the analysis of the output resistance in the saturation region. Only the first order term is kept in the Taylor series. For simplicity, we also assume that the contributions to the Early Voltage from all three mechanisms are independent and can be calculated separately.

2.6.1 Channel Length Modulation (CLM)

If channel length modulation is the only physical mechanism to be taken into account, then according to Eq. (2.6.3), the Early Voltage can be calculated by:

(2.6.4)

$$V_{ACLM} = I_{dsat} \left(\frac{\partial I_{ds}}{\partial L} \frac{\partial L}{\partial V_{ds}} \right)^{-1} = \frac{A_{bulk} E_{sat} L + V_{gst}}{A_{bulk} E_{sat}} \left(\frac{\partial \Delta L}{\partial V_{ds}} \right)^{-1}$$

where ΔL is the length of the velocity saturation region, the effective channel length is L-ΔL. Based on the quasi-two dimensional approximation, V_{ACLM} can be derived as the following:

(2.6.5)

$$V_{ACLM} = \frac{A_{bulk} E_{sat} L + V_{gst}}{A_{bulk} E_{sat} l} (V_{ds} - V_{dsat})$$

 BSIM3v3 Manual Copyright © 1995, 1996, UC Berkeley

where V_{ACLM} is the Early Voltage due to channel length modulation alone.

The parameter P_{clm} is introduced into the V_{ACLM} expression not only to compensate for the error caused by the Taylor expansion in the Early Voltage model, but also to compensate for the error in X_j since:

$$l \propto \sqrt{X_j}$$

and the junction depth, X_j, can not generally be determined very accurately. Thus, the V_{ACLM} became:

$$V_{ACLM} = \frac{1}{P_{clm}} \frac{A_{bulk} E_{sat} L + V_{gst}}{A_{bulk} E_{sat} l} (V_{ds} - V_{dsat}) \tag{2.6.6}$$

2.6.2 Drain-Induced Barrier Lowering (DIBL)

As discussed above, threshold voltage is can be approximated as a linear function of the drain voltage. According to Eq. (2.6.3), the Early Voltage due to the *DIBL* effect can be calculated as:

$$V_{ADIBLC} = I_{dsat} \left(\frac{\partial I_{ds}}{\partial V_{th}} \frac{\partial V_{th}}{\partial V_{ds}} \right)^{-1} \tag{2.6.7}$$

$$V_{ADIBLC} = \frac{(V_{gsteff} + 2v_t)}{\theta_{rout}(1 + P_{DIBLCB} V_{bseff})} \left(1 - \frac{A_{bulk} V_{dsat}}{A_{bulk} V_{dsat} + V_{gsteff} + 2v_t} \right)$$

During the derivation of Eq. (2.6.7), the parasitic resistance is assumed to be equal to 0. As expected, V_{ADIBLC} is a strong function of L as shown in Eq. (2.6.7). As channel length decreases, V_{ADIBLC} decreases very quickly.

The combination of the *CLM* and *DIBL* effects determines the output resistance in the third region, as was shown in Figure 2-6.

Despite the formulation of these two effects, accurate modeling of the output resistance in the saturation region requires that the coefficient $\theta_{th}(L)$ be replaced by $\theta_{rout}(L)$. Both $\theta_{th}(L)$ and $\theta_{rout}(L)$ have the same channel length dependencies, but different coefficients. The expression for $\theta_{rout}(L)$ *is:*

$$\theta_{rout}(L) = P_{diblc1}[\exp(-D_{rout}L/2l_t) + 2\exp(-D_{rout}L/l_t)] + P_{diblc2} \tag{2.6.8}$$

The variables P_{diblc1}, P_{diblc2}, P_{diblcb} and D_{rout} are the newly introduced parameters to correct for DIBL effect in the strong inversion region. The reason why D_{vt0} is not equal to P_{diblc1} and D_{vt1} is not equal to D_{rout} is because the gate voltage modulates the *DIBL* effect. When the threshold voltage is determined, the gate voltage is equal to the threshold voltage. But in the saturation region where the output resistance is modeled, the gate voltage is much larger than the threshold voltage. Drain induced barrier lowering may not be the same at different gate bias. P_{diblc2} is usually very small (may be as small as *8.0E-3*). If P_{diblc2} is placed into the threshold voltage model, it will not cause any significant change. However it is an important parameter in V_{ADIBL} for long channel devices, because P_{diblc2} will be dominant in Eq. (2.6.8) if the channel is long.

2.6.3 Current Expression without Substrate Current Induced

 BSIM3v3 Manual Copyright © 1995, 1996, UC Berkeley

Body Effect

In order to have a continuous drain current and output resistance expression at the transition point between linear and saturation region, the V_{Asat} parameter is introduced into the Early Voltage expression. V_{Asat} is the Early Voltage at $V_{ds} = V_{dsat}$ and is as follows:

(2.6.9)

$$V_{Asat} = \frac{E_{sat}L + V_{dsat} + 2R_{ds}v_{sat}C_{ox}W(V_{gst} - A_{bulk}V_{ds}/2)}{1 + A_{bulk}R_{ds}v_{sat}C_{ox}W}$$

Thus, the total Early Voltage, V_A, can be written as:

(2.6.10)

$$V_A = V_{Asat} + \left(\frac{1}{V_{ACLM}} + \frac{1}{V_{ADIBL}}\right)^{-1}$$

The complete (with no impact ionization at high drain voltages) current expression in the saturation region is given by:

(2.6.11)

$$I_{dso} = Wv_{sat}C_{ox}(V_{gst} - A_{bulk}V_{dsat})\left(1 + \frac{V_{ds} - V_{dsat}}{V_A}\right)$$

Furthermore, another parameter, P_{vag}, is introduced in V_A to account for the gate bias dependence of V_A more accurately. This much said, the final expression for Early Voltage becomes:

(2.6.12)

$$V_A = V_{Asat} + \left(1 + \frac{P_{vag}V_{gs}}{E_{sat}L_{eff}}\right)\left(\frac{1}{V_{ACLM}} + \frac{1}{V_{ADIBLC}}\right)^{-1}$$

2.6.4 Current Expression with Substrate Current Induced Body Effect

When the electrical field near the drain is very large (> 0.1MV/cm), some electrons coming from the source will be energetic (hot) enough to cause impact ionization. This creates electron-hole pairs when they collide with silicon atoms. The substrate current I_{sub} thus created during impact ionization will increase exponentially with the drain voltage. A well known I_{sub} model [20] is given as:

$$I_{sub} = \frac{A_i}{B_i} I_{dsat}(V_{ds} - V_{dsat}) \exp\left(-\frac{B_i l}{V_{ds} - V_{dsat}}\right) \tag{2.6.13}$$

The parameters A_i and B_i are determined from extraction. I_{sub} will affect the drain current in two ways. The total drain current will change because it is the sum of the channel current from the source as well as the substrate current. The total drain current can now be expressed [21] as follows:

$$\begin{aligned} I_{ds} &= I_{dso} + I_{sub} \\ &= I_{dso}\left[1 + \frac{(V_{ds} - V_{dsat})}{\frac{B_i}{A_i}\exp\left(\frac{B_i l}{V_{ds} - V_{dsat}}\right)}\right] \end{aligned} \tag{2.6.14}$$

The total drain current, including *CLM, DIBL* and *SCBE*, can be written as:

$$I_{ds} = W v_{sat} C_{ox}(V_{gst} - A_{bulk} V_{dsat})\left(1 + \frac{V_{ds} - V_{dsat}}{V_A}\right)\left(1 + \frac{V_{ds} - V_{dsat}}{V_{ASCBE}}\right) \tag{2.6.15}$$

 BSIM3v3 Manual Copyright © 1995, 1996, UC Berkeley

where V_{ASCBE} can also be called as the Early Voltage due to the substrate current induced body effect. Its expression is the following:

$$V_{ASCBE} = \frac{B_i}{A_i}\exp\left(\frac{B_i l}{V_{ds} - V_{dsat}}\right) \tag{2.6.16}$$

From Eq. (2.6.16), we can see that V_{ASCBE} is a strong function of V_{ds}. In addition, we also observe that V_{ASCBE} is small only when V_{ds} is large. This is why *SCBE* is important for devices with high drain voltage bias. The channel length and gate oxide dependence of V_{ASCBE} comes from V_{dsat} and l. In BSIM3v3, we replace *Bi* with *PSCBE2* and *Ai/Bi* with *PSCBE1/L* to yield the following expression for V_{ASCBE}:

$$\frac{1}{V_{ASCBE}} = \frac{P_{SCBE2}}{L}\exp\left(-\frac{P_{SCBE1} l}{V_{ds} - V_{dsat}}\right) \tag{2.6.17}$$

The variables P_{scbe1} and P_{scbe2} are determined experimentally.

2.7 Subthreshold Drain Current

The drain current equation in the subthreshold region was given in [2, 3] can is expressed by the following:

$$I_{ds} = I_{s0}\left(1 - \exp\left(-\frac{V_{ds}}{v_t}\right)\right)\exp\left(\frac{V_{gs} - V_{th} - V_{off}}{n v_t}\right) \tag{2.7.1}$$

(2.7.2)

$$I_{s0} = \mu_0 \frac{W}{L} \sqrt{\frac{q\varepsilon_{si} N_{ch}}{2\phi_s}} v_t^2$$

Here the parameter *vt* is the thermal voltage and is given by KBT/q. Voff is the offset voltage, as discussed in Jeng's dissertation [18]. Voff is an important parameter which determines the drain current at $V_{gs} = 0$. In Eq. (2.7.1), the parameter *n* is the subthreshold swing parameter. Experimental data shows that the subthreshold swing is a function of channel length and the interface state density. These two mechanisms are modeled in BSIM3v3 by the following:

(2.7.3)

$$n = 1 + N_{factor} \frac{C_d}{C_{ox}} + \frac{(C_{dsc} + C_{dscd} V_{ds} + C_{dscb} V_{bseff}) \left(\exp(-D_{VT1} \frac{L_{eff}}{2 l_t}) + 2 \exp(-D_{VT1} \frac{L_{eff}}{l_t}) \right)}{C_{ox}} + \frac{C_{it}}{C_{ox}}$$

where, the term:

$$(C_{dsc} + C_{dscd} V_{ds} + C_{dscb} V_{bseff}) \left(\exp(-D_{VT1} \frac{L_{eff}}{2 l_t}) + 2 \exp(-D_{VT1} \frac{L_{eff}}{l_t}) \right)$$

represents the coupling capacitance between the drain or source to the channel. The parameters Cdsc, Cdscd, Cdscb are extracted. The parameter C_{it} in Eq. (2.7.3) is the capacitance due to interface states. From Eq. (2.7.3), it can be seen that subthreshold swing shares the same exponential dependence on channel length as the *DIBL* effect. The parameter *Nfactor* is introduced to compensate for errors in the depletion width capacitance calculation. *Nfactor* is determined experimentally and is usually very close to 1.

2.8 Effective Channel Length and Width:

The effective channel length and width used in all model expressions is given below:

(2.8.1)

$$L_{eff} = L_{drawn} - 2dL$$

(2.8.2a)

$$W_{eff} = W_{drawn} - 2dW$$

(2.8.2b)

$$W_{eff}' = W_{drawn} - 2dW'$$

The only difference between Eq. (2.8.1a) and (2.8.1b) is that the former includes bias dependencies. The parameters dW and dL are modeled by the following:

(2.8.3)

$$dW = dW' + dW_g V_{gsteff} + dW_b \left(\sqrt{\phi_s - V_{bseff}} - \sqrt{\phi_s}\right)$$

$$dW' = W_{\text{int}} + \frac{W_l}{L^{W\ln}} + \frac{W_w}{W^{Wwn}} + \frac{W_{wl}}{L^{W\ln} W^{Wwn}}$$

(2.8.4)

$$dL = L_{\text{int}} + \frac{L_l}{L^{L\ln}} + \frac{L_w}{W^{Lwn}} + \frac{L_{wl}}{L^{L\ln} W^{Lwn}}$$

These complicated formulations require some explanation. From Eq. (2.8.3), the variable W_{int} models represents the tradition manner from which "delta W" is extracted (from the *int*ercepts of straights lines on a 1/R_{ds} vs. W_{drawn} plot). The parameters dW_g and dW_b have been added to account for the contribution of both front gate and back side (substrate) biasing effects. For dL, the parameter L_{int}

represents the traditional manner from which "delta L" is extracted (mainly from the *int*ercepts of lines from a R_{ds} vs. L_{drawn} plot).

The remaining terms in both *dW* and *dL* are included for the convenience of the user. They are meant to allow the user to model each parameter as a function of W(drawn), L(drawn), and their associated product terms. In addition, the freedom to model these dependencies as other than just simple inverse functions of W and L is also provide for the user in BSIM3v3. For *dW*, they are *Wln* and *Wwn*. For *dL* they are *Lln* and *Lwn*.

By default all of the above *geometrical* dependencies for both dW and dL are turned off. Again, these equations are provided in BSIM3v3 only for the convenience of the user. As such, it is up to the user must adopt the correct extraction strategy to ensure proper use.

2.9 Poly Gate Depletion Effect

When a gate voltage is applied to a heavily doped poly-silicon gate, e.g. NMOS with n+ poly-silicon gate, a thin depletion layer will be formed at the interface between the poly-silicon and gate oxide. Although this depletion layer is very thin due to the high doping concentration of the poly-Si gate, its effect cannot be ignored in the 0.1µm regime since the gate oxide thickness will also be very small, possibly 50Å or thinner.

Figure 2-7 shows a NMOSFET with a depletion region in the n+ poly-silicon gate. The doping concentration in the n+ poly-silicon gate is N_{gate} and the doping concentration in the substrate is N_{sub}. The gate oxide thickness is T_{ox}. The depletion width in the poly gate is X_p. The depletion width in the substrate is X_d. If we

 BSIM3v3 Manual Copyright © 1995, 1996, UC Berkeley

assume the doping concentration in the gate is infinite, then no depletion region will exist in the gate, and there would be one sheet of positive charge whose thickness is zero at the interface between the poly-silicon gate and gate oxide.

In reality, the doping concentration is, of course, finite. The positive charge near the interface of the poly-silicon gate and the gate oxide is distributed over a finite depletion region with thickness X_p. In the presence of the depletion region, the voltage drop across the gate oxide and the substrate will be reduced, because part of the gate voltage will be dropped across the depletion region in the gate. That means the effective gate voltage will be reduced.

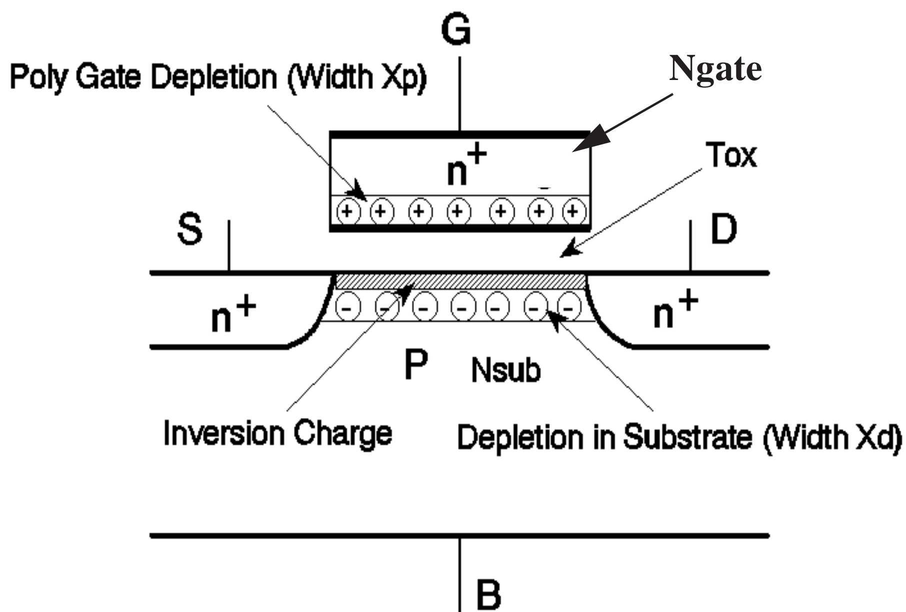

Figure 2-7. Charge distribution in a MOSFET with the poly gate depletion effect. The device is in the strong inversion region.

The effective gate voltage can be calculated in the following manner. Assume the doping concentration in the poly gate is uniform. The voltage drop in the poly gate (V_{poly}) can be calculated as:

$$V_{poly} = \frac{1}{2} X_{poly} E_{poly} = \frac{qN_{gate} X_{poly}^2}{2\varepsilon_{si}} \tag{2.9.1}$$

where E_{poly} is the maximum electrical field in the poly gate. The boundary condition at the interface of poly gate and the gate oxide is

$$\varepsilon_{ox} E_{ox} = \varepsilon_{si} E_{poly} = \sqrt{2q\varepsilon_{si} N_{gate} V_{poly}} \tag{2.9.2}$$

where E_{ox} is the electrical field in the gate oxide. The gate voltage satisfies

$$V_{gs} - V_{FB} - \phi_s = V_{poly} + V_{ox} \tag{2.9.3}$$

where V_{ox} is the voltage drop across the gate oxide and satisfies $V_{ox} = E_{ox} T_{ox}$.

According to the equations (2.9.1) to (2.9.3), we obtain the following:

$$a(V_{gs} - V_{FB} - \phi_s - V_{poly})^2 - V_{poly} = 0 \tag{2.9.4}$$

where

$$a = \frac{\varepsilon_{ox}^2}{2q\varepsilon_{si} N_{gate} T_{ox}^2} \tag{2.9.5}$$

 BSIM3v3 Manual Copyright © 1995, 1996, UC Berkeley

By solving the equation (2.9.4), we get the effective gate voltage (V_{gs_eff}) which is equal to:

$$V_{gs_eff} = V_{FB} + \phi_s + \frac{q\varepsilon_{si} N_{gate} T_{ox}^2}{\varepsilon_{ox}^2}\left(\sqrt{1 + \frac{2\varepsilon_{ox}^2 (V_{gs} - V_{FB} - \phi_s)}{q\varepsilon_{si} N_{gate} T_{ox}^2}} - 1\right) \tag{2.9.6}$$

Figure 2-8 shows V_{gs_eff} / V_{gs} versus the gate voltage. The threshold voltage is assumed to be *0.4V*. If T_{ox} = *40 Å*, the effective gate voltage can be reduced by *6%* due to the poly gate depletion effect as the applied gate voltage is equal to *3.5V*.

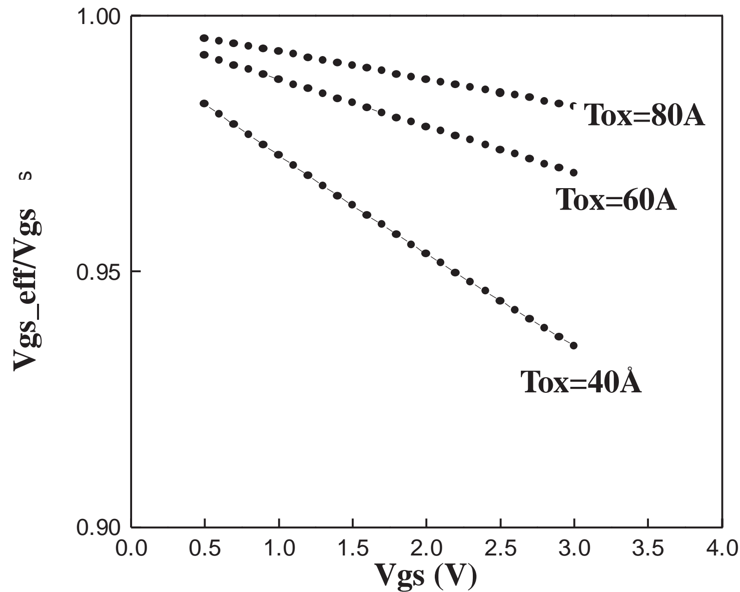

Figure 2-8. The effective gate voltage versus applied gate voltage at different gate oxide thickness.

The drain current reduction in the linear region as a function of the gate voltage can now be determined. Assume the drain voltage is very small, e.g. *50mV*. Then

the linear drain current is proportional to $C_{ox}(V_{gs} - V_{th})$. The ratio of the linear drain current with and without poly gate depletion is equal to:

$$\frac{I_{ds}(V_{gs_eff})}{I_{ds}(V_{gs})} = \frac{(V_{gs_eff} - V_{th})}{(V_{gs} - V_{th})} \tag{2.9.7}$$

Figure 2-9 shows Ids(Vgs_eff) / Ids(Vgs) versus the gate voltage using eq. (2.9.7). The drain current can be reduced by several percent due to gate depletion.

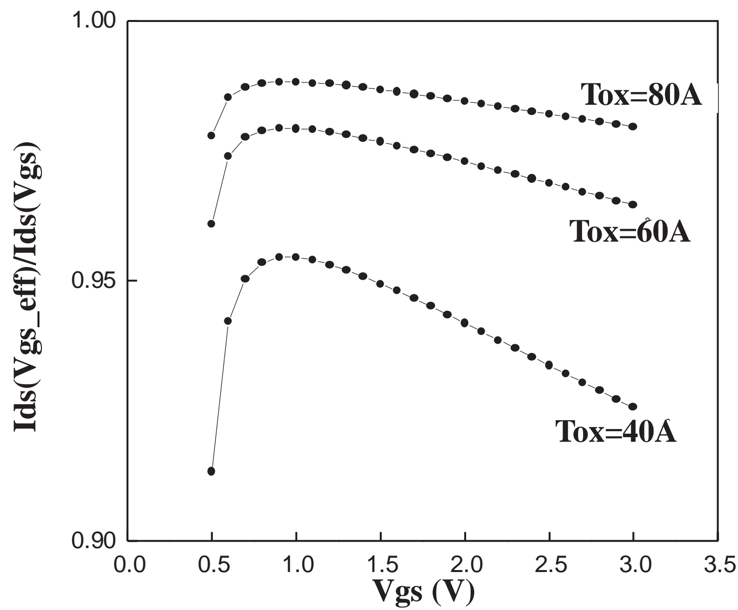

Figure 2-9. Ratio of linear region current with poly gate depletion effect and that without.