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

The development of BSIM3v3 is based on Poisson's equation using gradual channel approximation and coherent quasi 2D analysis, taking into account the effects of device geometry and process parameters. BSIM3v3.2.2 considers the following physical phenomena observed in MOSFET 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.**
- **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 ($V_{th}$) is one of the most important requirements for precise description of device 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 and MOSFET is operating in the strong inversion region and drift current is dominant. Second, if the gate voltage is smaller than $V_{th}$, the inversion charge density is smaller than the substrate doping concentration. The transistor is 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 $V_{th}$, the inversion charge density is close to the doping concentration and the MOSFET is operating in the transition region. In such a case, diffusion and drift currents are both important.

For MOSFET's with long channel length/width and uniform substrate doping concentration, $V_{th}$ is given by [2]:

$$V_{th} = V_{FB} + \Phi_s + g\sqrt{\Phi_s - V_{bs}} = V_{Tideal} + g\left(\sqrt{\Phi_s - V_{bs}} - \sqrt{\Phi_s}\right) \tag{2.1.1}$$

where $V_{FB}$ is the flat band voltage, $V_{Tideal}$ is the threshold voltage of the long channel device at zero substrate bias, and $\gamma$ is the body bias coefficient and is given by:

$$g = \frac{\sqrt{2e_{si}qN_a}}{C_{ox}} \tag{2.1.2}$$

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

$$\Phi_s = 2\frac{k_B T}{q}\ln\left(\frac{N_a}{n_i}\right) \tag{2.1.3}$$

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. Modifications have to be made when the substrate doping concentration is not uniform and/or when the channel length is short, narrow, or both.

### 2.1.1 Vertical Non-Uniform Doping Effect

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

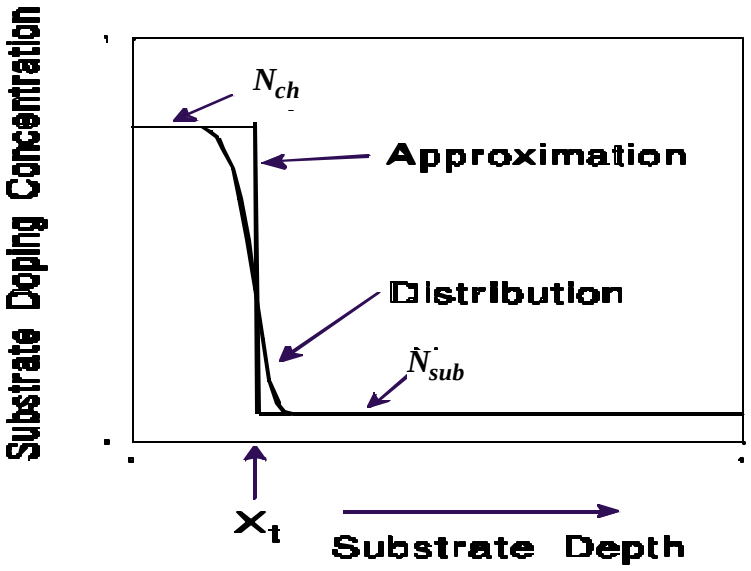


**Figure 2-1. Actual substrate doping distribution and its approximation.**

The substrate doping concentration is usually higher near the $Si/SiO_2$ interface (due to $V_{th}$ adjustment) than deep into the 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 $\gamma$ 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 profile, the following $V_{th}$ model is proposed:

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

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:

$$K_1 = g_2 - 2K_2\sqrt{\Phi_s - V_{bm}} \tag{2.1.5}$$

$$K_2 = \frac{(g_1 - g_2)\left(\sqrt{\Phi_s - V_{bx}} - \sqrt{\Phi_s}\right)}{2\sqrt{\Phi_s}\left(\sqrt{\Phi_s - V_{bm}} - \sqrt{\Phi_s}\right) + V_{bm}} \tag{2.1.6}$$

where $\gamma_1$ and $\gamma_2$ are body bias coefficients when the substrate doping concentration are equal to $N_{ch}$ and $N_{sub}$, respectively:

$$g_1 = \frac{\sqrt{2qe_{si}N_{ch}}}{C_{ox}} \tag{2.1.7}$$

$$g_2 = \frac{\sqrt{2qe_{si}N_{sub}}}{C_{ox}} \tag{2.1.8}$$

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

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

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. $N_{ch}$, $N_{sub}$ and $X_t$). 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 source/drain 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 $V_{th}$ to increase in magnitude because the average doping concentration in the channel is larger. The average channel doping concentration can be calculated as follows:

(2.1.10)

$$N_{eff} = \frac{N_a(L-2L_x)+Npocket2L_x}{L} = N_a\left(1+\frac{2L_x}{L}\cdot\frac{Npocket-N_a}{N_a}\right)$$
$$\equiv N_a\cdot\left(1+\frac{Nlx}{L}\right)$$

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

(2.1.11)

$$V_{th} = V_{th0} + K_1\left(\sqrt{\Phi_s - V_{bs}} - \sqrt{\Phi_s}\right) - K_2V_{bs} + K_1\left(\sqrt{1+\frac{Nlx}{L_{eff}}} - 1\right)\sqrt{\Phi_s}$$

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 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].

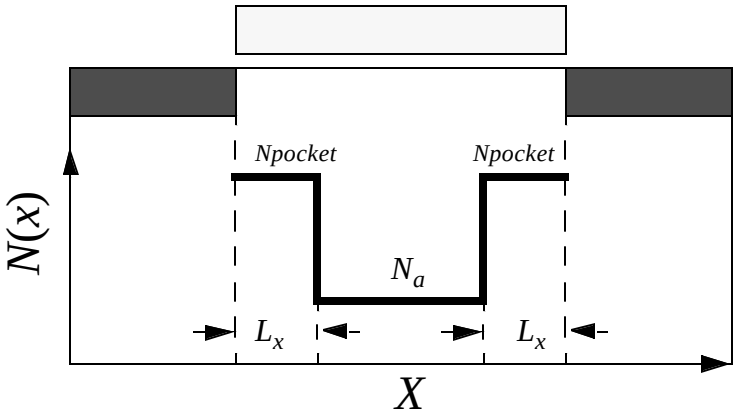


**Figure 2-2. Lateral doping profile is non-uniform.**

### 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, 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. The short-channel effect is included in the $V_{th}$ model as:

$$V_{th} = V_{th0} + K_1\left(\sqrt{\Phi_s - V_{bs}} - \sqrt{\Phi_s}\right) - K_2 V_{bs} + K_1\left(\sqrt{1+\frac{Nlx}{L_{eff}}} - 1\right)\sqrt{\Phi_s} - \Delta V_{th} \tag{2.1.12}$$

where $\Delta V_{th}$ is the threshold voltage reduction due to the short channel effect. Many models have been developed to calculate $\Delta V_{th}$. 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]. This model was derived by solving the quasi 2D Poisson equation along the channel. This quasi-2D model concluded that:

$$\Delta V_{th} = q_{th}(L)\left(2\left(V_{bi} - \Phi_s\right) + V_{ds}\right) \tag{2.1.13}$$

where $V_{bi}$ is the built-in voltage of the PN junction between the source and the substrate and 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 is the source/drain doping concentration with a typical value of around $1 \times 10^{20}$ cm$^{-3}$. The expression $\theta_{th}(L)$ is a short channel effect coefficient, which has a strong dependence on the channel length and is given by:

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

$l_t$ is referred to as the *characteristic length* and is given by

$$l_t = \sqrt{\frac{e_{si} T_{ox} X_{dep}}{e_{ox} h}} \tag{2.1.16}$$

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

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

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

Based on the above discussion, the influences of drain/source charge sharing and *DIBL* effects on$V_{th}$ are described by (2.1.15). In order to make the model fit different technologies, several parameters such as $D_{vt0}$, $D_{vt2}$,

*Dsub*, *Eta*0 and *Etab* are introduced, and the following modes are used to account for charge sharing and *DIBL* effects separately.

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

$$\Delta V_{th}(L) = q_{th}(L)(V_{bi} - \Phi_s) \tag{2.1.19}$$

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

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

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

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 ($N_{ch}$). $D_{vt0}$, $D_{vt1}$,$D_{vt2}$, *Eta*0, *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. Thus 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, $\Delta V_{th}$ 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 MOSFET's; therefore the threshold voltage reduction due to the short channel effect will be smaller in *LDD*-MOSFET's.

As the body bias becomes more negative, the depletion width will increase as shown in Eq. (2.1.17). Hence $\Delta V_{th}$ 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 $\Delta V_{th}$ 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's 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 an 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.

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

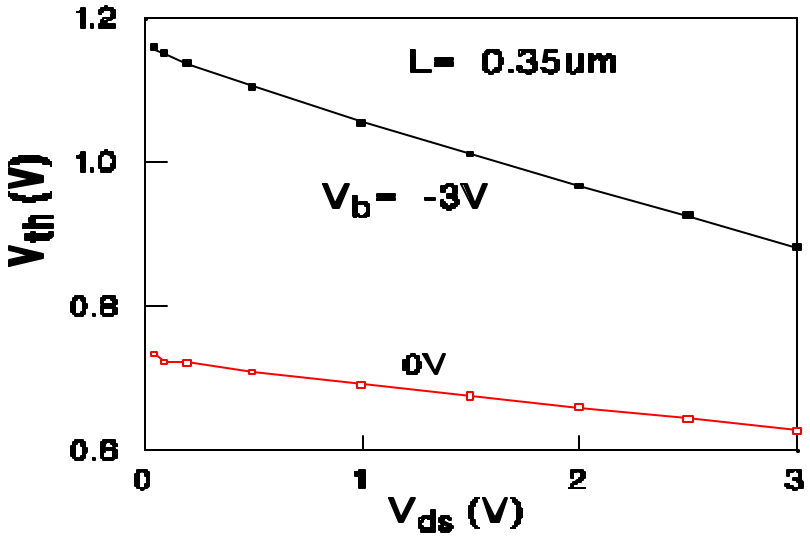


**Figure 2-3. Threshold voltage versus the drain voltage at different body biases.**

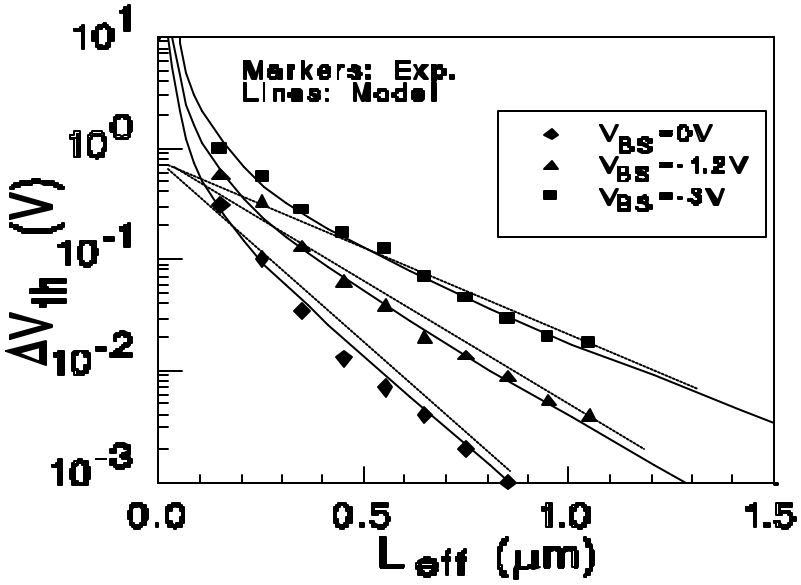


**Figure 2-4. Channel length dependence of threshold voltage.**

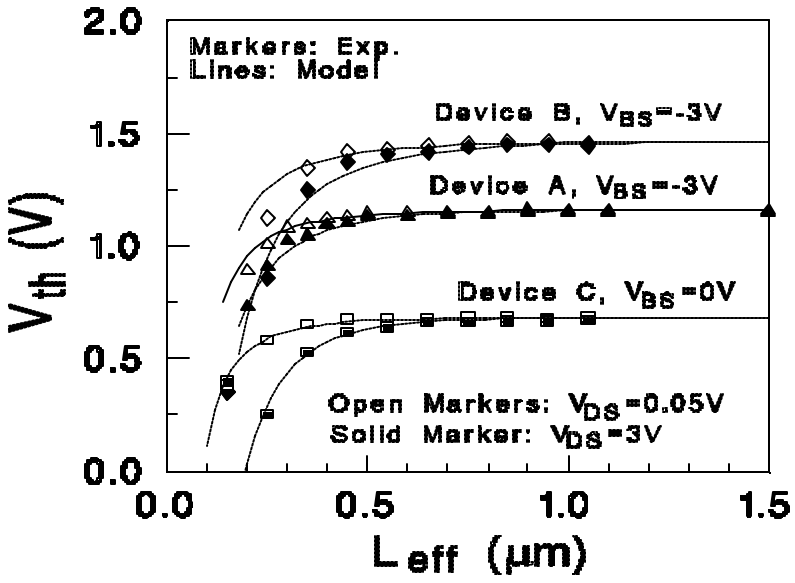


**Figure 2-5. Threshold voltage versus channel length at different biases.**

### 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 in $V_{th}$. It is shown in [2] that this increase can be modeled as:

(2.1.23)

$$\frac{pqN_a X_{d\max}^2}{2C_{ox}W} = 3p\,\frac{T_{ox}}{W}\,\Phi_s$$

 

The right hand side of Eq. (2.1.23) represents the additional voltage increase. This change in $V_{th}$ is modeled 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, parameters $K_3$, $K_{3b}$, and $W_0$ are introduced as

(2.1.24a)

$$\left(K_3 + K_{3b}V_{bs}\right)\frac{T_{ox}}{W_{eff}'+W_0}\Phi_s$$

$W_{eff}$' is the effective channel width (with no bias dependencies), which will be defined in Section 2.8. In addition, we must consider the narrow width effect for small channel lengths. To do this we introduce the following:

(2.1.24b)

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

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

(2.1.25)

$$
\begin{aligned}
V_{th} &= V_{th0ox} + K_{1ox}\cdot\sqrt{\Phi_s - V_{bseff}} - K_{2ox}V_{bseff} \\
&+ K_{1ox}\left(\sqrt{1+\frac{Nlx}{L_{eff}}}-1\right)\sqrt{\Phi_s} + \left(K_3 + K_{3b}V_{bseff}\right)\frac{T_{ox}}{W_{eff}'+W_0}\Phi_s \\
&- D_{VT0w}\left(\exp\left(-D_{VT1w}\frac{W_{eff}'L_{eff}}{2l_{tw}}\right)+2\exp\left(-D_{VT1w}\frac{W_{eff}'L_{eff}}{l_{tw}}\right)\right)\left(V_{bi}-\Phi_s\right) \\
&- D_{VT0}\left(\exp\left(-D_{VT1}\frac{L_{eff}}{2l_t}\right)+2\exp\left(-D_{VT1}\frac{L_{eff}}{l_t}\right)\right)\left(V_{bi}-\Phi_s\right) \\
&- \left(\exp\left(-D_{sub}\frac{L_{eff}}{2l_{to}}\right)+2\exp\left(-D_{sub}\frac{L_{eff}}{l_{to}}\right)\right)\left(E_{tao}+E_{tab}V_{bseff}\right)V_{ds}
\end{aligned}
$$

where $T_{ox}$ dependence is introduced in the model parameters $K_1$ and $K_2$ to improve the scalibility of $V_{th}$ model with respect to $T_{ox}$. $V_{th0ox}$, $K_{1ox}$ and $K_{2ox}$ are modeled as

$$V_{th0ox} = V_{th0} - K_1\cdot\sqrt{\Phi_s}$$

and

$$K_{1ox} = K_1\cdot\frac{T_{ox}}{T_{oxm}}$$

$$K_{2ox} = K_2\cdot\frac{T_{ox}}{T_{oxm}}$$

$T_{oxm}$ is the gate oxide thickness at which parameters are extracted with a default value of $T_{ox}$.

In Eq. (2.1.25), all $V_{bs}$ terms have been substituted with a $V_{bseff}$ 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 since unreasonable values can occur if this expression is not introduced (see Section 3.8 for details).

$$V_{bseff} = V_{bc} + 0.5[V_{bs} - V_{bc} - d_1 + \sqrt{(V_{bs} - V_{bc} - d_1)^2 - 4d_1V_{bc}}\ ] \tag{2.1.26}$$

where $\delta_1$ = 0.001V. The parameter $V_{bc}$ 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, and is equal to:

$$V_{bc} = 0.9\left(\Phi_s - \frac{K_1^{\ 2}}{4K_2^{\ 2}}\right)$$

# 2.2 Mobility Model

A good mobility model is critical to the accuracy of a MOSFET model. The scattering mechanisms responsible for surface mobility basically include phonons, coulombic scattering, 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 the gate oxide thickness, substrate doping concentration, threshold voltage, gate and substrate voltages, 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

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

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 given by

$$m_{eff} = \frac{m_0}{1+(E_{eff}/E_0)^n} \tag{2.2.2}$$

Values for $\mu_0$, $E_0$, and $\nu$ 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) |
|---|---|---|
| $\mu_0$ ($cm^2/Vsec$) | 670 | 160 |
| $E_0$ ($MV/cm$) | 0.67 | 0.7 |
| $\nu$ | 1.6 | 1.0 |

**Table 2-1. Typical mobility values for electrons and holes.**

For an 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 very time consuming power function in SPICE simulation. Taylor expansion Eq. (2.2.2) is used, and the coefficients are left to be determined by experimental data or to be obtained by fitting the unified formulation. Thus, we have

(mobMod=1) (2.2.4)

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

where $V_{gst}=V_{gs}-V_{th}$. To account for depletion mode devices, another mobility model option is given by the following

(mobMod=2) (2.2.5)

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

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)

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

## 2.3 Carrier Drift Velocity

Carrier drift velocity is also one of the most important parameters. The following velocity saturation equation [17] is used in the model

(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)

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

## 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 non-uniform along the channel length. This will cause $V_{th}$ to vary along the channel. This effect is called bulk charge effect [14].

The parameter, $A_{bulk}$, is used to take into account the bulk charge effect. Several extracted parameters such as $A_0$, $B_0$, $B_1$ 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 bulk charge effect under high 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 is given by

(2.4.1)

$$A_{bulk} = \left(1 + \frac{K_{1ox}}{2\sqrt{\Phi_s - V_{bseff}}}\left(\frac{A_0 L_{eff}}{L_{eff} + 2\sqrt{X_J X_{dep}}}\left(1 - A_{gs} V_{gsteff}\left(\frac{L_{eff}}{L_{eff} + 2\sqrt{X_J X_{dep}}}\right)^2\right) + \frac{B_0}{W_{eff}' + B_1}\right)\right) \cdot \frac{1}{1 + Keta V_{bseff}}$$

where $A_0$, $A_{gs}$, $B_0$, $B_1$ and *Keta* are determined by experimental data. Eq. (2.4.1) shows that $A_{bulk}$ is very close to unity 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 ($R_{ds}$=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*.

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

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

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. (2.5.2) from $y = 0$ to $y = L$ and $V(y) = 0$ to $V(y) = V_{ds}$, 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 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 region can now be 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 ($R_{ds}$>0)

Parasitic source/drain resistance is an important device parameter which can affect MOSFET performance significantly. As channel length scales down, the parasitic resistance will not be proportionally scaled. As a result, $R_{ds}$ will have a more significant impact on device characteristics. Modeling of parasitic resistance in a direct method yields a complicated drain current expression. In order to make simulations more efficient, the parasitic resistances is modeled such that the resulting drain current equation in the linear region can be calculateed [3] as

(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 the 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). $V_{dsat}$ with parasitic resistance $R_{ds}$ becomes

(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)

$$a = A_{bulk}^2 R_{ds} C_{ox} W v_{sat} + (\frac{1}{l} - 1) A_{bulk}$$

$$b = -(V_{gst}(\frac{2}{l} - 1) + A_{bulk} E_{sat} L + 3A_{bulk} R_{ds} C_{ox} W v_{sat} V_{gst})$$

$$c = E_{sat} L V_{gst} + 2R_{ds} C_{ox} W v_{sat} V_{gst}^2$$

$$l = A_1 V_{gst} + A_2$$

The last expression for $\lambda$ is introduced to account for non-saturation effect of the device. The parasitic resistance is modeled as:

(2.5.11)

$$R_{ds} = \frac{R_{dsw}\left(1 + P_{rwg} V_{gsteff} + P_{rwb}\left(\sqrt{\Phi_s - V_{bseff}} - \sqrt{\Phi_s}\right)\right)}{\left(10^6 W_{eff}'\right)^{W_r}}$$

The variable $R_{dsw}$ is the resistance per unit width, $W_r$ is a fitting parameter, $P_{rwb}$ and $P_{rwg}$ are the body bias and the gate bias coeffecients, repectively.

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

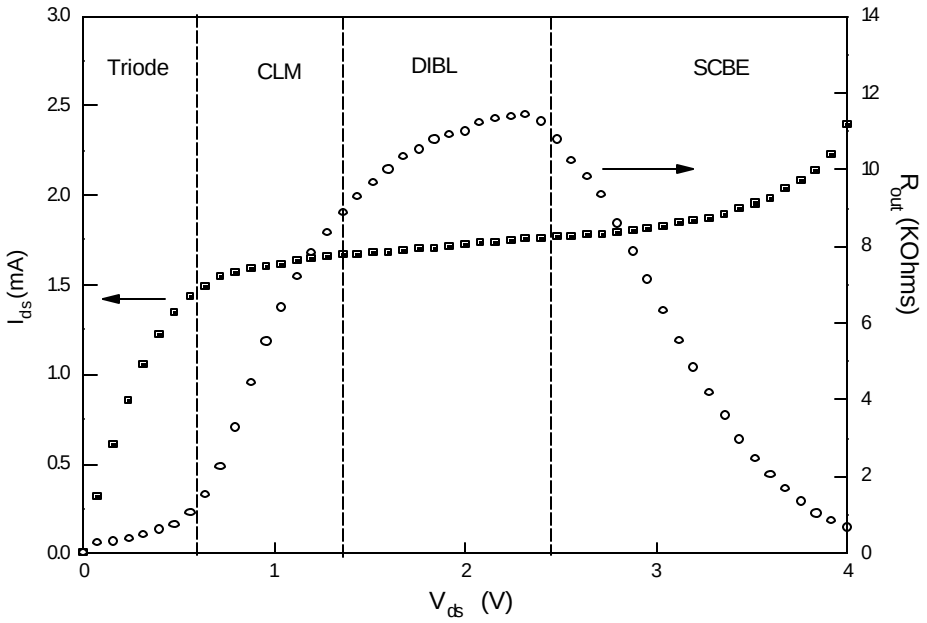


**Figure 2-6. General behavior of MOSFET output resistance.**

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].

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

where

$$I_{dsat} = I_{ds}(V_{gs,} V_{dsat}) = W v_{sat} C_{ox}(V_{gst} - A_{bulk} V_{dsat}) \tag{2.6.2}$$

and

$$V_A = I_{dsat}(\frac{\partial I_{ds}}{\partial V_{ds}})^{-1} \tag{2.6.3}$$

The parameter $V_A$ is called the Early voltage 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. 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

$$V_{ACLM} = I_{dsat}(\frac{\P I_{ds}}{\P L}\frac{\P L}{\P V_{ds}})^{-1} = \frac{A_{bulk}E_{sat}L + V_{gst}}{A_{bulk}E_{sat}}(\frac{\P DL}{\P V_{ds}})^{-1} \tag{2.6.4}$$

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

$$V_{ACLM} = \frac{A_{bulk}E_{sat}L + V_{gst}}{A_{bulk}E_{sat}l}(V_{ds} - V_{dsat}) \tag{2.6.5}$$

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 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:

(2.6.7)

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

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

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

(2.6.8)

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

Parameters $P_{diblc1}$, $P_{diblc2}$, $P_{diblcb}$ and $D_{rout}$ are introduced 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 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:

$$V_{Asat} = \frac{E_{sat}L + V_{dsat} + 2R_{ds}v_{sat}C_{ox}W(V_{gst} - A_{bulk}V_{ds}/2)}{1 + A_{bulk}R_{ds}v_{sat}C_{ox}W} \tag{2.6.9}$$

Total Early voltage, $V_A$, can be written as

$$V_A = V_{Asat} + \left(\frac{1}{V_{ACLM}} + \frac{1}{V_{ADIBL}}\right)^{-1} \tag{2.6.10}$$

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

$$I_{dso} = Wv_{sat} C_{ox} (V_{gst} - A_{bulk} V_{dsat})(1 + \frac{V_{ds} - V_{dsat}}{V_A}) \tag{2.6.11}$$

Furthermore, another parameter, $P_{vag}$, is introduced in $V_A$ to account for the gate bias dependence of $V_A$ more accurately. The final expression for Early voltage becomes

$$V_A = V_{Asat} + (1 + \frac{P_{vag} V_{gs}}{E_{sat} L_{eff}})(\frac{1}{V_{ACLM}} + \frac{1}{V_{ADIBLC}})^{-1} \tag{2.6.12}$$

### 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:

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

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

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

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

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

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

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

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.* We replace *Bi* with *PSCBE*2 and *Ai/Bi* with *PSCBE*1*/L* to yield the following expression for $V_{ASCBE}$

(2.6.17)

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

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

# 2.7 Subthreshold Drain Current

The drain current equation in the subthreshold region can be expressed as [2, 3]

(2.7.1)

$$I_{ds} = I_{s0}(1-\exp(-\frac{V_{ds}}{v_t}))\exp(\frac{V_{gs}-V_{th}-V_{off}}{nv_t})$$

(2.7.2)

$$I_{s0} = m_0\frac{W}{L}\sqrt{\frac{qe_{si}N_{ch}}{2f_s}}v_t^2$$

Here the parameter $v_t$ is the thermal voltage and is given by $K_BT/q$. $V_{off}$ is the offset voltage, as discussed in Jeng's dissertation [18]. $V_{off}$ 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 by the following

(2.7.3)

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

where the term

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

represents the coupling capacitance between the drain or source to the channel. The parameters $C_{dsc}$, $C_{dscd}$ and $C_{dscb}$ 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

$$L_{eff} = L_{drawn} - 2dL \tag{2.8.1}$$

$$W_{eff} = W_{drawn} - 2dW \tag{2.8.2a}$$

$$W_{eff}' = W_{drawn} - 2dW' \tag{2.8.2b}$$

The only difference between Eq. (2.8.2a) and (2.8.2b) 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 provided for the user. 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 for the convenience of the user. As such, it is up to the user to 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 an 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 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.

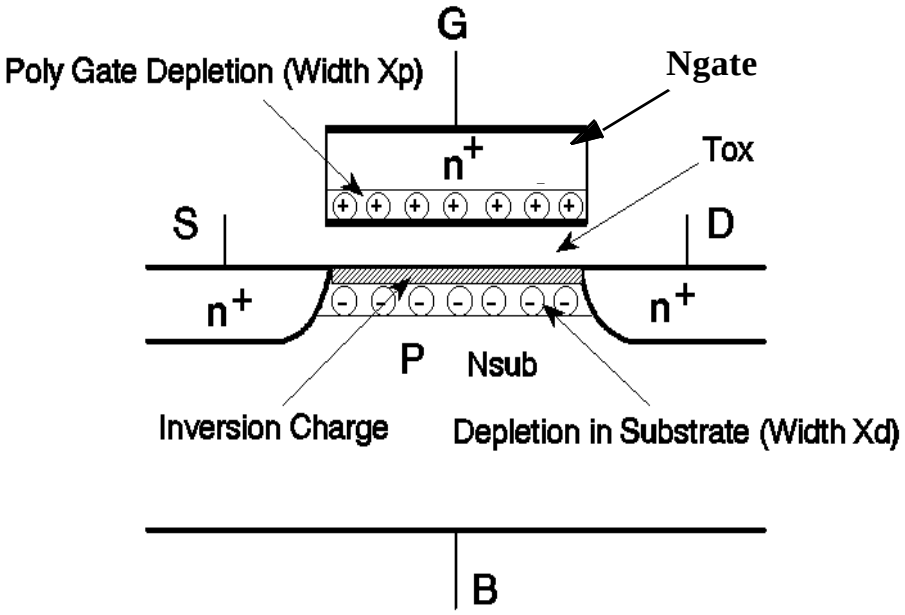


**Figure 2-7. Charge distribution in a MOSFET with the poly gate depletion effect. The device is in the strong inversion region.**

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

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

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

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

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

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

where (2.9.5)

$$a = \frac{e_{ox}^{\ 2}}{2qe_{si}N_{gate}T_{ox}^{\ 2}}$$

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

$$V_{gs\_eff} = V_{FB} + \Phi_s + \frac{qe_{si}N_{gate}T_{ox}^{\ 2}}{e_{ox}^{\ 2}}\left(\sqrt{1 + \frac{2e_{ox}^{\ 2}\left(V_{gs} - V_{FB} - \Phi_s\right)}{qe_{si}N_{gate}T_{ox}^{\ 2}}} - 1\right) \tag{2.9.6}$$

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.

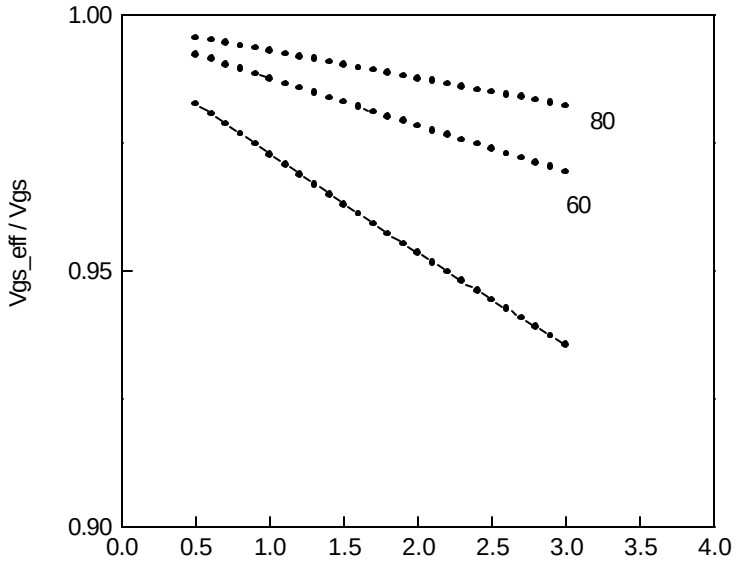


**Figure 2-8. The effective gate voltage versus applied gate voltage at different gate oxide thickness.**

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:

(2.9.7)

Figure 2-9 shows $I_{ds}(V_{gs\_eff})$ / $I_{ds}(V_{gs})$ versus the gate voltage using Eq. (2.9.7). The drain current can be reduced by several percent due to gate depletion.