Journal of Circuits, Systems, and Computers, Vol. 11, No. 6 (2002) 1–26
© World Scientific Publishing Company

LEAKAGE CURRENT IN DEEP-SUBMICRON CMOS CIRCUITS

KAUSHIK ROY,* SAIBAL MUKHOPADHYAY† and HAMID MAHMOODI-MEIMAND‡

School of Electrical and Computer Engineering, Purdue University,
West Lafayette, IN, USA
**kaushik@ecn.purdue.edu*
†sm@ecn.purdue.edu
‡mahmoodi@ecn.purdue.edu

The high leakage current in deep submicron regimes is becoming a significant contributor to the power dissipation of CMOS circuits as the threshold voltage, channel length, and gate oxide thickness are reduced. Consequently, the identification and modeling of different leakage components is very important for the estimation and reduction of leakage power, especially in the low power applications. This paper explores the various transistor intrinsic leakage mechanisms including the weak inversion, the drain-induced barrier lowering, the gate-induced drain leakage, and the gate oxide tunneling.

Keywords: ???

1. Introduction

To achieve higher density and performance, and lower power consumption, MOS devices have been scaled for more than 30 years. Transistor delay times have decreased by more than 30% per technology generation resulting in the doubling of microprocessor performance in every two years. Supply voltage ($V_{\rm DD}$) has been scaling down at the rate of 30% per technology generation in order to keep power consumption under control. Hence, the transistor threshold voltage ($V_{\rm th}$) has to be commensurately scaled to maintain high drive current and achieve performance improvement of at least 30% per technology generation. However, the threshold voltage scaling results in the substantial increase of the subthreshold leakage current.[1]

Transistor off-state current ($I_{\rm OFF}$) is the drain current when the gate-to-source voltage is zero. $I_{\rm OFF}$ is influenced by the threshold voltage, channel physical dimensions, channel/surface doping profile, drain/source junction depth, gate oxide thickness and $V_{\rm DD}$. $I_{\rm OFF}$ in long channel devices is dominated by the leakage from the drain-well and well-substrate reverse bias p-n junctions.[2] Short channel transistors require lower power supply levels to reduce internal electric fields and power consumption. This forces a reduction in the threshold voltage, $V_{\rm th}$, that causes a relatively large increase in $I_{\rm OFF}$. This increase is due to the weak inversion state leakage which is a function of $V_{\rm th}$ and is not due to transistor channel length. In this paper we focus on all leakage mechanisms contributing to standby leakage (not just the drain terminal). Other leakage mechanisms are peculiar to the small geometries

themselves. As drain voltage V_{D} increases, the drain-to-channel depletion region widens and significant drain current can result. This increase in I_{OFF} is typically due to the channel surface current from drain-induced barrier lowering (DIBL) or due to the deep channel punchthrough currents.[3–7] Moreover, as the channel width decreases, both the threshold voltage and the off current get modulated by the width of the transistor, giving rise to significant narrow-width effects. To maintain a reasonable short-channel effect immunity while scaling down the channel length, the oxide thickness has to be decreased below 20 A^0, for CMOS devices beyond the 100 nm node. A decrease in oxide thickness results in an increase in electric field across the gate oxide. The high electric field and low oxide thickness result in considerable current flowing through the gate of a transistor. This current destroys the classical infinite input impedance assumption of MOS transistors and thus affects the circuit performance severely. Major contributors to the gate leakage current are the gate oxide tunneling and the injection of hot carrier from substrate to gate oxide. Gate induce drain leakage (GIDL) is another significant leakage mechanism, resulting from the depletion at the drain surface below the gate-drain overlap region. Figure 1 shows the projections for some transistor physical dimensions, supply voltage and device power consumption according to the International Technology Roadmap for Semiconductors.[8] All the parameters are normalized to their values in the year 2001. As shown in Fig. 1(b), due to the substantial increase in leakage current, the static power consumption is expected to exceed switching component of power consumption unless effective measures are taken to reduce leakage power.

Due to the short channel effects, the channel length cannot be arbitrarily reduced even if allowed by lithography. For digital applications, the most undesirable short channel effect is the reduced gate threshold voltage at which the device turns on, especially at high drain voltages. Therefore to take the best advantage of the new high-resolution lithographic techniques, new device designs, structures, and technologies should be developed to keep short channel effect under control at very small dimensions. In addition to the gate oxide thickness and junction scaling, another technique to improve short channel characteristics is well engineering. By changing the doping profile in the channel region, the distribution of the electric field and potential contours can be changed. The goal is to optimize the channel profile to minimize the off-state leakage while maximizing the linear and saturated drive currents. Super Steep Retrograde Wells (SSRW) and halo implants have been used as a means to scale the channel length and increase the transistor drive current without causing an increase in the off-state leakage current.[9–12]

In this paper different leakage current components and mechanisms in deep submicron transistors are explored which is essential to guide solutions for reducing power and leakage per transistor.

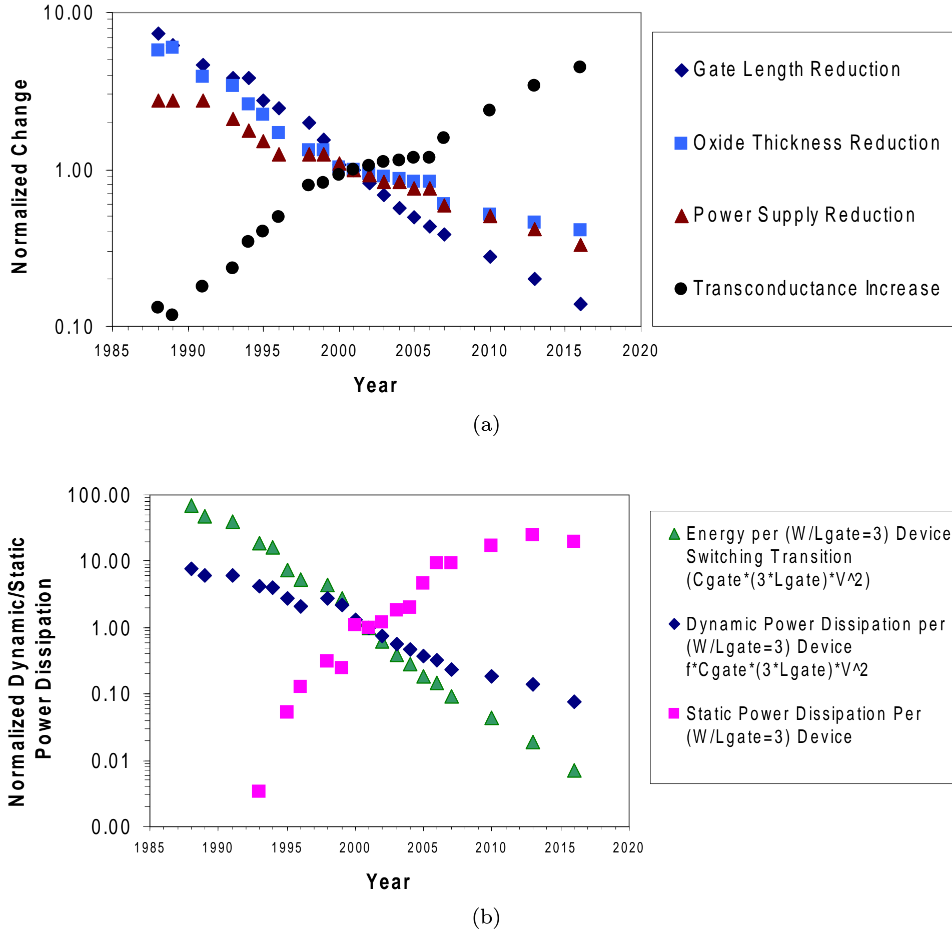

Fig. 1. ITRS projection for transistor scaling trends and power consumption (a) physical dimensions and supply voltage (b) device power consumption.[8]

2. Transistor Leakage Mechanisms

Figure 2 shows a typical $\log(I_{\rm D})$ versus $V_{\rm G}$ curve. It allows measurement of many device parameters such as $I_{\rm OFF}$, $V_{\rm th}$, $I_{\rm D}$(SAT), $I_{\rm D}$(LIN), g_m(SAT), g_m(LIN), and slope (S) of $V_{\rm G}$ versus $I_{\rm D}$ in the weak inversion state. $I_{\rm OFF}$ is measured at the $V_{\rm G} = 0$ V intercept. The n-channel transistor in Fig. 2 has an $I_{\rm OFF}$ of 20 pA/μm and 4 pA/μm in the saturated and linear states. We describe six short channel leakage mechanisms as illustrated in Fig. 3. I_1 is reverse bias p-n junction leakage, I_2 is the subthreshold leakage, I_3 is the oxide leakage, and I_4 is the gate current due to hot carrier injection. I_5 is the gate induced drain leakage (GIDL), and I_6 is the channel punch-through. Currents I_1, I_2, I_5, I_6 are off-state leakage mechanisms while I_3 (oxide tunneling) occurs in both ON and OFF states. I_4 can occur in the off-state, but more typically occurs during the transistor bias states in transition.

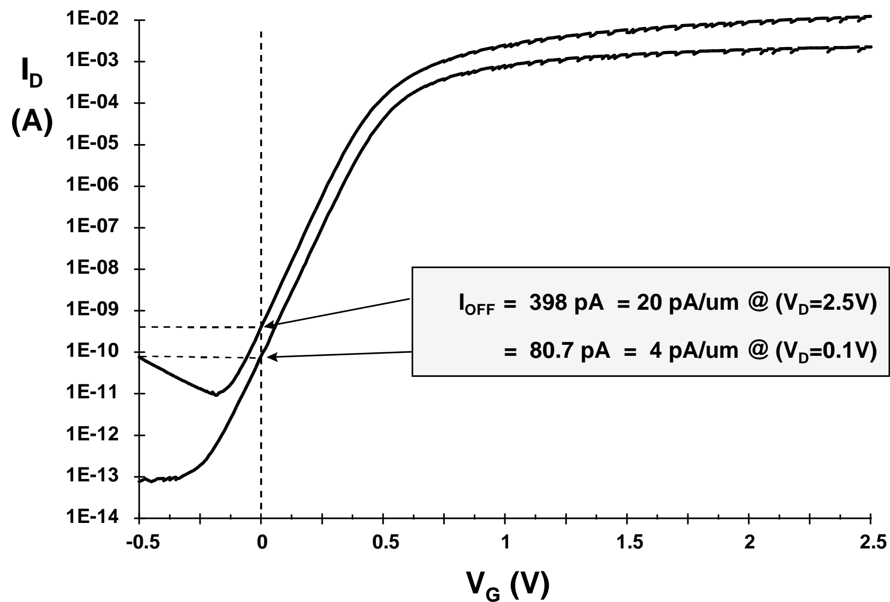

Fig. 2. Log (I_{D}) versus V_{G} at saturated bias ($V_{\mathrm{D}} = 2.5$ V) and linear bias ($V_{\mathrm{D}} = 0.1$ V) states for 20×0.4 μm n-channel transistor.[29]

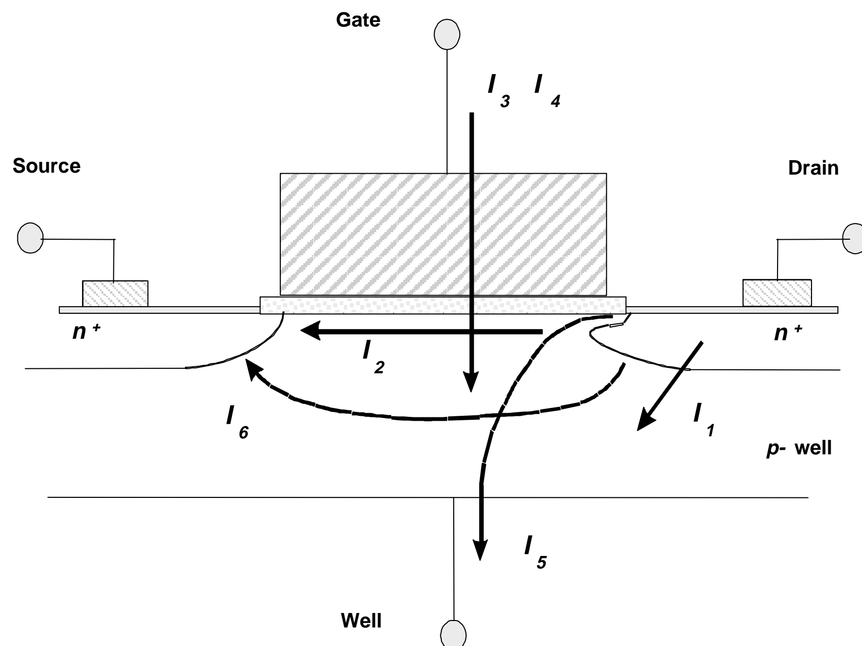

Fig. 3. Summary of leakage current mechanisms of deep submicron transistors.

2.1. *p-n junction reverse bias current (and gated diode leakage)* (I_1)

A reverse bias *p-n* junction leakage (I_1) has two main components: One is the minority carrier diffusion/drift near the edge of the depletion region and the other is due to the electron-hole pair generation in the depletion region of the reverse bias junction.[6] If both *n*- and *p*-regions are heavily doped (this will be the case for advanced MOSFETs using heavily doped shallow junctions and halo doping for better SCE), Zener and band-to-band tunneling may also be present. For an MOS transistor, additional leakage can occur between the drain and well junction from gated diode device action (overlap and vicinity of gate to the drain-to-well *p-n* junctions) or carrier generation in drain-to-well depletion regions with influences of the gate on these current components.[13] *p-n* junction reverse bias leakage ($I_{\rm REV}$) is a function of junction area and doping concentration.[5,6]

2.2. *Subthreshold leakage* (I_2)

Subthreshold or weak inversion conduction current between source and drain in a MOS transistor occurs when gate voltage is below $V_{\rm th}$.[5,14] The weak inversion region is seen in Fig. 2 as the linear portion of the curve. In the weak inversion, the minority carrier concentration is small but not zero. Figure 4 shows the variation of minority carrier concentration along the length of the channel. Let us consider that the source of the *n*-channel MOSFET is grounded, $V_g < V_{\rm th}$, and the drain-to-source voltage $|V_{\rm ds}| \geq 0.1$ V. In conditions where weak inversion occurs, $V_{\rm ds}$ drops

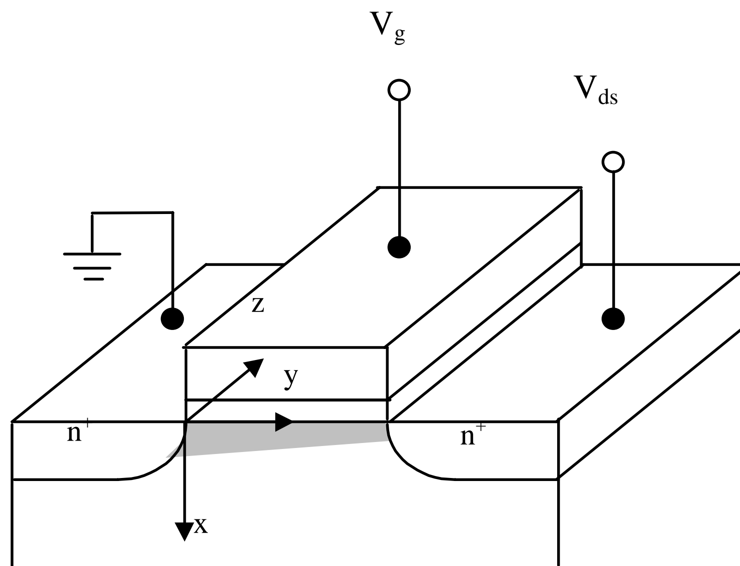

Fig. 4. Variation of minority carrier concentration in the channel of a MOSFET biased in weak inversion.

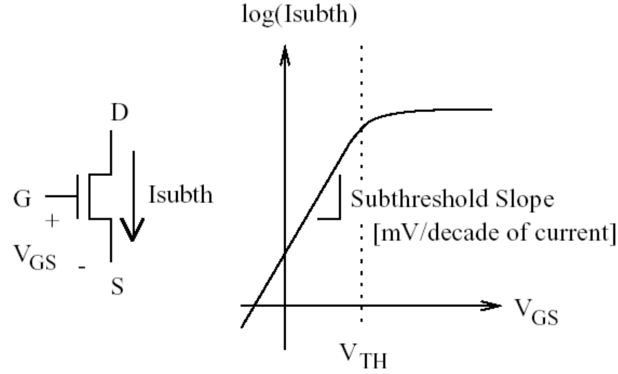

Fig. 5. Subthreshold leakage in an NMOS transistor.

almost entirely across the reverse-biased substrate-drain *p-n* junction. As a result, the variation along the channel (the y axis) in the electrostatic potential ϕ_s at the semiconductor surface is small. The y component, $\mathbf{E}_y$ of the electric field vector $\mathbf{E}$, being equal to $\partial\phi/\partial y$, is also small. With both the number of mobile carriers and the longitudinal electric field small, the drift component of the subthreshold drain-to-source current is negligible. Therefore, unlike the strong inversion region in which the drift current dominates, subthreshold conduction is dominated by the diffusion current. The carriers move by diffusion along the surface similar to charge transport across the base of bipolar transistors. The exponential relation between the driving voltage on the gate and the drain current is a straight line in a semi log plot (Fig. 5). Weak inversion typically dominates modern device off-state leakage due to the low V_{th} used.

The weak inversion current can be expressed based on the following equation.[14]

$$I_{\text{ds}} = \mu_0 C_{\text{ox}} \frac{W}{L}(m-1)(v_{\text{T}})^2 \times e^{\frac{(V_g - V_{\text{th}})}{m v_{\text{T}}}} \times (1 - e^{\frac{-v_{\text{DS}}}{v_{\text{T}}}}), \tag{1}$$

where

$$m = 1 + \frac{C_{\text{dm}}}{C_{\text{ox}}} = 1 + \frac{\frac{\varepsilon_{\text{si}}}{W_{\text{dm}}}}{\frac{\varepsilon_{\text{ox}}}{t_{\text{ox}}}} = 1 + \frac{3t_{\text{ox}}}{W_{\text{dm}}}\,, \tag{2}$$

V_{th} is the threshold voltage and $v_{\text{T}} = KT/q$ is the thermal voltage. C_{ox} is the gate oxide capacitance, μ_0 is the zero bias mobility and m is the subthreshold swing coefficient (also called body effect coefficient) for the transistor. W_{dm} is the maximum depletion layer width and t_{ox} is the gate oxide thickness. C_{dm} is the capacitance of the depletion layer and C_{ox} is the capacitance of the insulator layer.

In long channel devices, the subthreshold current is independent of the drain voltage for V_{DS} larger than few v_{T}. On the other hand, the dependency on the gate voltage is exponential as illustrated in Fig. 5. The inverse of the slope of the

$\log_{10}(I_{\text{ds}})$ versus V_{gs} characteristic is called the subthreshold slope (S_t).[14]

$$S_t = \left(\frac{d(\log_{10} I_{\text{ds}})}{dV_g}\right)^{-1} = 2.3\frac{mkT}{q} = 2.3\frac{kT}{q}\left(1 + \frac{C_{\text{dm}}}{C_{\text{ox}}}\right). \quad (3)$$

The subthreshold slope indicates how effectively the flow of the drain current of a device can be stopped when V_{gs} is decreased below V_{th}. As the device dimensions and the supply voltage are being scaled down to enhance performance, power efficiency, and reliability, this characteristic becomes a limitation on how small a power supply can be used. The parameter S_t is measured in milivolts per decade. For the limiting case of $t_{\text{ox}} \to 0$ and at room temperature, $S_t \approx 60$ mV/decade. Typical S_t values for a bulk CMOS process can range from 80 mV/decade to 120 mV/decade or more. A low value for subthreshold slope is most desirable. It can be noted from the above expression that S_t can be made smaller by using a thinner oxide (insulator) layer to reduce t_{ox} or a lower substrate doping concentration (resulting in larger W_{dm}). Changes in operating conditions, namely lower temperature or a substrate bias, also causes S_t to decrease.

2.2.1. *Drain-induced barrier lowering*

In long-channel devices, the source and drain are separated far enough such that their depletion regions have no effect on the potential or field pattern in most part of the device, and hence, the threshold voltage is virtually independent of the channel length and drain bias. In a short-channel device, however, the source and drain depletion width in the vertical direction, and the source-drain potential has a strong effect on the band bending over a significant portion of the device. Therefore, the threshold voltage and consequently the subthreshold current of short-channel devices vary with the drain bias. This effect is referred to as drain-induced barrier lowering (DIBL). One way to describe it is to consider the energy barrier at the surface between the source and drain, as shown in Fig. 6.[14] Under OFF conditions, this energy barrier prevents electrons from flowing to the drain. For a long-channel device, the barrier height is mainly controlled by the gate voltage and is not sensitive to V_{ds}. However, the barrier of a short-channel device reduces along with the increase of drain voltage, which causes a higher subthreshold current and lower threshold voltage.

DIBL occurs when the depletion region of the drain interacts with the source near the channel surface to lower the source potential barrier. It happens when a high drain voltage is applied to a short-channel device, lowering the barrier height and resulting in further decrease of the threshold voltage. The source then injects carriers into the channel surface without the gate playing a role. DIBL is enhanced at a higher drain voltage and shorter L_{eff}. Surface DIBL typically happens before deep bulk punchthrough. Ideally, DIBL does not change the slope, S_t, but it does lower V_{th}. Higher surface and channel doping and shallow source/drain junction depths reduce the DIBL effect on the subthreshold leakage current.[14,15] Figure 7

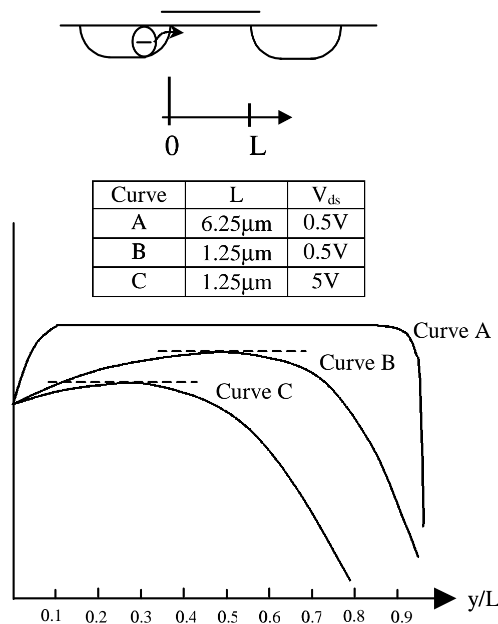

Fig. 6. Lateral energy band diagram at the surface versus distance (normalized to the channel length L) from the source to the drain for (a) long-channel MOSFET, (b) a short-channel MOSFET, (c) a short channel MOSFET at high drain bias. The gate voltage is the same for all three cases.[14]

illustrates the DIBL effect as it moves the curve up and to the left as V_{D} increases. DIBL can be measured at constant V_{G} as the change in I_{D} corresponds to a change in V_{D}.

2.2.2. *Body effect*

Reverse biasing well to source junction of a MOSFET transistor widens the bulk depletion region and increases the threshold voltage, V_{th}.[5,6] The effect of body bias can be considered in the threshold voltage equation[14]:

$$V_{\mathrm{th}} = V_{\mathrm{fb}} + 2\psi_{\mathrm{B}} + \frac{\sqrt{2\varepsilon_{\mathrm{st}} q N_a (2\psi_{\mathrm{B}} + V_{sb})}}{C_{\mathrm{ox}}}, \tag{4}$$

where V_{fb} is the flat band voltage, N_{a} is doping density in the substrate, and $\psi_{\mathrm{B}} = (KT/q)\ln(N_a/n_i)$ is the difference between the Fermi potential and the intrinsic potential in the substrate. The slope of V_{th} versus V_{sb} curve is therefore,

$$\frac{dV_{\mathrm{th}}}{dV_{\mathrm{bs}}} = \frac{\sqrt{\varepsilon_{\mathrm{st}} q N_a / 2(2\psi_{\mathrm{B}} + V_{\mathrm{sb}})}}{C_{\mathrm{ox}}}, \tag{5}$$

which is referred to as the substrate sensitivity. It can be seen from Eq. (5) that the substrate sensitivity is higher for higher bulk doping concentration and the substrate sensitivity decreases as the substrate reverse bias increases. At $V_{\mathrm{sb}} = 0$,

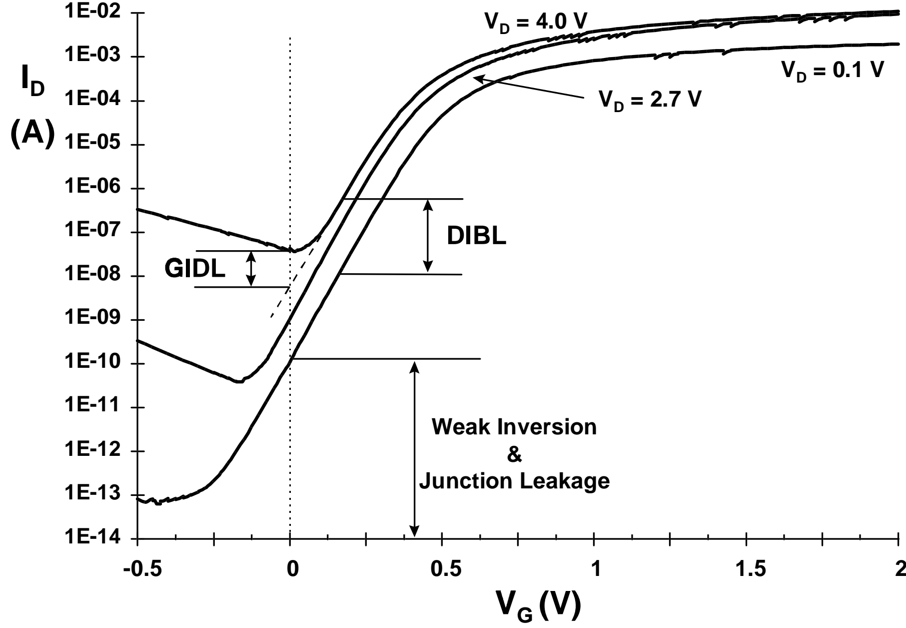

Fig. 7. n-channel $I_{\rm D}$ versus $V_{\rm G}$ showing DIBL, GIDL, weak inversion, and p-n junction reverse bias leakage components.[29]

the substrate sensitivity is $C_{\rm dm}/C_{\rm ox}$ or $m-1$ according to Eq. (2). Therefore, m is also called the body effect coefficient.

Figure 8 shows suppression in the n-channel drain current when the well-to-source voltage is back biased from 0 to -5 V (the back bias is the well voltage). Virtually no change is seen in the subthreshold slope S_t (Fig. 8) in contrast to the temperature effect (Fig. 14). An important observation from Fig. 8 is that as $V_{\rm th}$ increases because of applied reverse substrate bias and due to a shift in I–V, $I_{\rm OFF}$ decreases.

The subthreshold leakage of a MOS device including weak inversion, DIBL, and body effect, can be modeled according to the following equation.[16]

$$I_{\rm subth} = A \times e^{\frac{1}{mv_{\rm T}}(V_{\rm G}-V_{\rm S}-V_{\rm th0}-\gamma'\times V_{\rm S}+\eta\times V_{\rm DS})} \times (1 - e^{\frac{-v_{\rm DS}}{V_{\rm T}}})\,, \tag{6}$$

where

$$A + \mu_0 C'_{\rm ox}\frac{W}{L_{\rm eff}}(v_{\rm T})^2 e^{1.8} e^{\frac{-\Delta V_{\rm th}}{\eta v_{\rm T}}}\,. \tag{7}$$

$V_{\rm th0}$ is the zero bias threshold voltage, and $v_{\rm T} = KT/q$ is the thermal voltage. The body effect for small values of source to bulk voltages is very nearly linear and is represented by the term $\gamma' V_{\rm S}$, where γ' is the linearized body effect coefficient. η is the DIBL coefficient, $C_{\rm ox}$ is the gate oxide capacitance, μ_0 is the zero bias mobility and m is the subthreshold swing coefficient for the transistor. $\Delta V_{\rm th}$ is a term introduced to account for the transistor-to-transistor leakage variations.

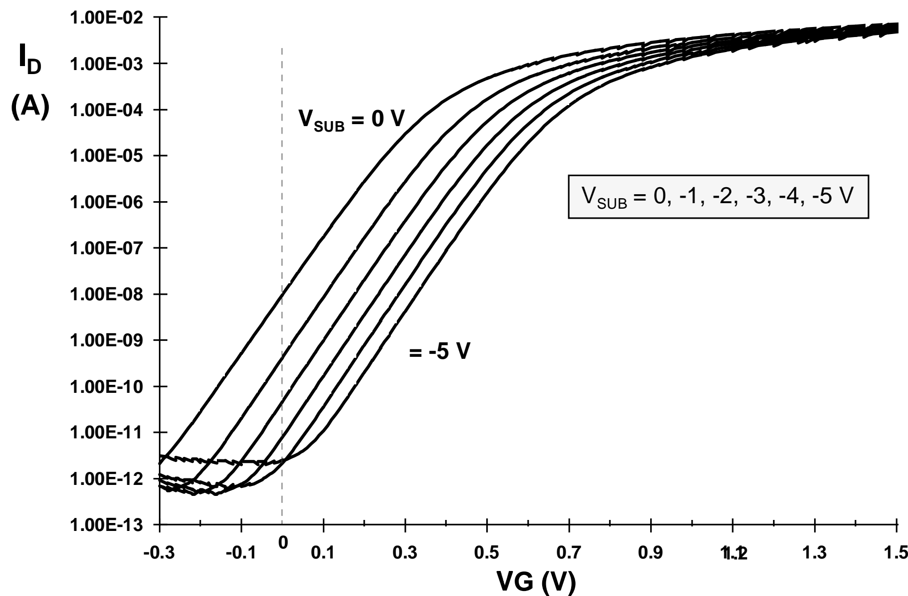

Fig. 8. n-channel $\log(I_D)$ versus V_G for six substrate biases on a 0.35 μm logic process technology ($V_D = 2.7$ V).[29]

2.2.3. *Narrow width effect*

The decrease of gate width modulates the threshold voltage of the transistor and thereby modulating the subthreshold leakage. There are mostly three narrow-width effects, which modulate the threshold voltage. The first effect in the case of local oxide isolation (LOCOS) gate MOSFET is the existence of the fringing field that causes the spreading of the gate induced depletion region to the exterior of the defined channel width and under the isolations as shown in Fig. 9(b). This results in the increase of total depletion charge in the bulk region above its otherwise expected value. The threshold voltage of MOS can be defined using depletion approximation as[17]

$$V_{th} = V_{fb} + \phi_s + \frac{Q_B}{C_{ox}} \tag{8}$$

where

$$\begin{aligned} V_{fb} &= \text{flat band voltage}\,, \\ \phi_s &= \text{surface potential}\,, \\ C_{ox} &= \text{capacitance across oxide}\,, \\ Q_B &= \text{depletion charge in the bulk}\,. \end{aligned}$$

Due to the narrow channel effect, Q_B increases by ΔQ_B as shown in Fig. 9(b). This results in an increase of the threshold voltage. This effect becomes more substantial as the channel width decreases and the depletion region underneath the fringing

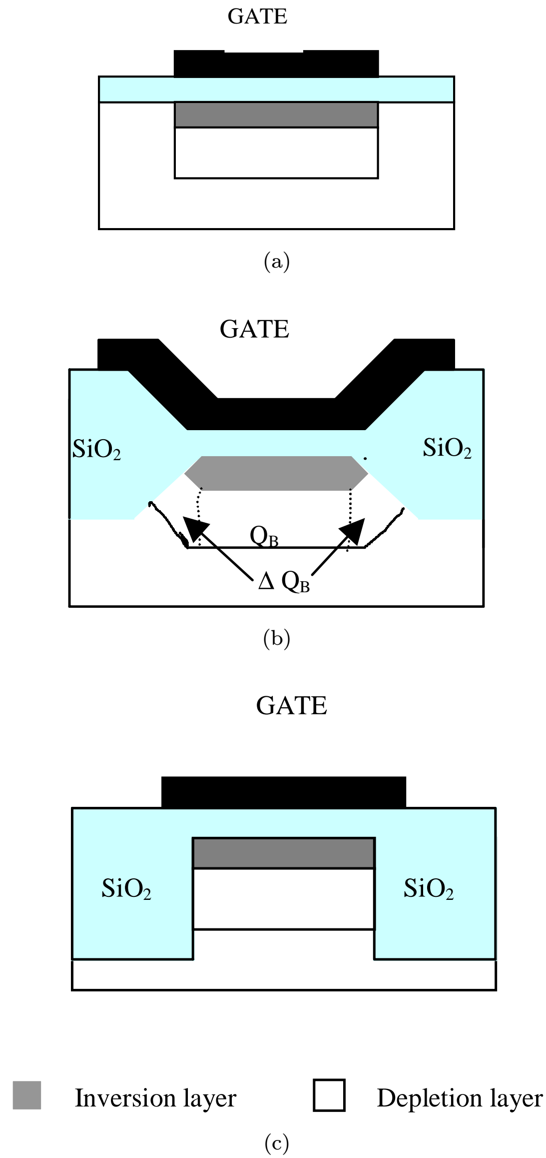

Fig. 9. Three types of device structure and associated inversion-depletion layer (a) large-geometry MOSFET, (b) LOCOS gate MOSFET, (c) Trench isolated MOSFET.[26]

field becomes comparable to the classical depletion formed by the vertical field. This results in the increase of the threshold voltage due to the narrow channel effect.[18,19] This narrow width effect can be modeled as an increase in $V_{\rm th}$ by an amount[18]

$$V_{\rm NCE} = \frac{\pi q N_{\rm sub} x_{\rm d,max}^2}{2C_{\rm ox} W_{\rm eff}} = 3\pi \frac{t_{\rm ox}}{W_{\rm eff}} \phi_{\rm S}\,, \tag{9}$$

where

$$N_{\rm sub} = \text{substrate doping}\,,$$
$$x_{\rm d,max} = \text{maximum vertical depletion width}\,,$$
$$C_{\rm ox} = \text{capacitance across oxide}\,,$$
$$W_{\rm eff} = \text{effective width}\,,$$
$$t_{\rm ox} = \text{oxide thickness}\,,$$
$$\phi_{\rm s} = \text{surface potential}\,.$$

The more accurate modeling can be found in Ref. 19.

The second narrow-width factor in case of LOCOS gate arises from the fact that channel doping is higher along the width dimension, due to the channel stop dopants encroaching under the gate. Hence, a higher voltage is needed to completely invert the channel.[20]

A more complex effect is seen in trench isolation devices, which is known as inverse-narrow width effect. In case of trench isolation devices, depletion layer cannot spread under the oxide isolation (Fig. 9(c)), hence reducing the possibility of an increase in the total depletion charge in the bulk and an increase in threshold voltage. On the other hand, due to the two-dimensional field induced edge-fringing effect at the gate edge, the formation of inversion layer at the edges occurs at a lower voltage than required at the center. Also, the overall gate capacitance ($C_{\rm T}$) now includes the sidewall capacitance ($C_{\rm F}$) due to the overall gate width with isolation oxide, hence increasing the overall gate capacitance.[17] Overall gate capacitance

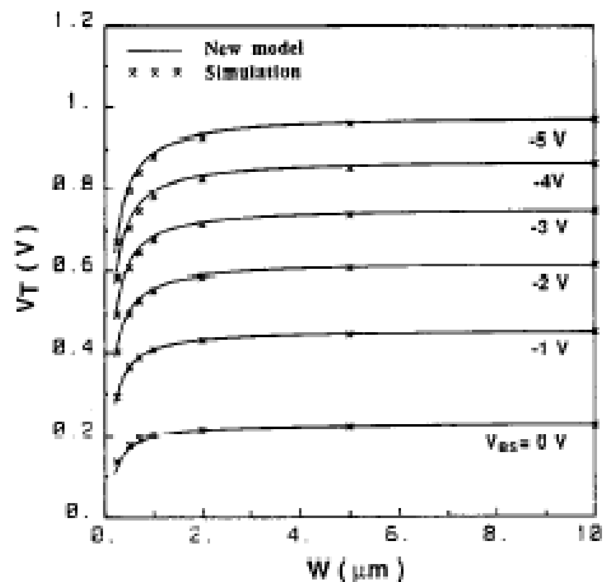

Fig. 10. Variation of threshold voltage with gate width for uniform doping using the model introduced in Ref. 17.

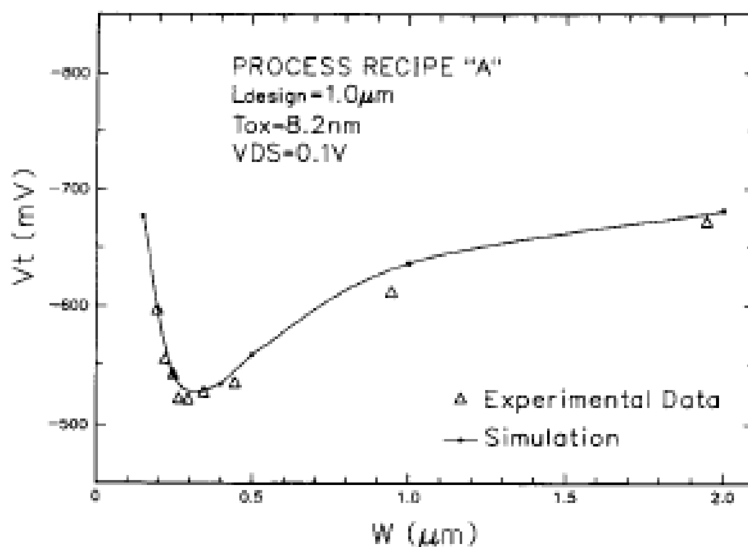

Fig. 11. Variation of threshold voltage with gate width in case of trench isolated buried channel p-MOSFET showing the anomalous behavior.[21]

is now given by $C_{\rm T} = C_{\rm ox}W + 2C_{\rm F}$, which is greater than $C_{\rm ox}$ in Eq. (8). Hence, overall $V_{\rm th}$ reduces. Figure 10 explains the behavior with the model introduced in Ref. 17.

A much more complex behavior can be observed in the case of trench-isolated buried channel p-MOSFETs, where reduction of width first decreases the $V_{\rm th}$ till the width is 0.4 μm, and thereafter a sharp increase in $V_{\rm th}$ is observed (Fig. 11). A more detailed description of the behavior is described in Ref. 21.

2.2.4. *Effect of channel length and $V_{\rm th}$ roll-off*

Threshold voltage of MOSFET decreases as the channel length is reduced. This reduction of threshold voltage with the reduction of channel length is known as the $V_{\rm th}$ roll-off. Figure 12 shows the reduction of threshold voltage with reduction in channel length. The principal reason behind this effect is the presence of two-dimensional field patterns in short-channel device instead of a one-dimensional field pattern in a long-channel device. This two-dimensional field pattern originates from the proximity of source drain region.[14] There are depletion regions surrounding the source-drain junctions. In a long channel device, since the source and drain are far apart, their depletion region does not have much effect on the potential profile or field pattern in most parts of the channel. However, in the case of short channel devices, the source drain distance is comparable to the depletion width in the vertical direction. As a result, source drain depletion width has a more pronounce effect. The source and drain depletion region now penetrates more into the channel length, resulting in the depletion in a part of the channel. Thus, the gate voltage

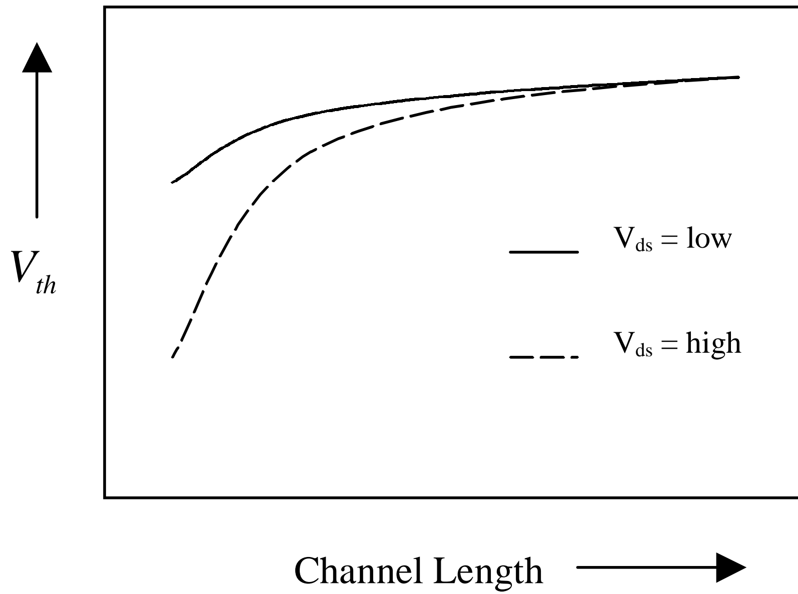

Fig. 12. Threshold voltage roll off with change in channel length; rate of decrease is more with higher drain bias.

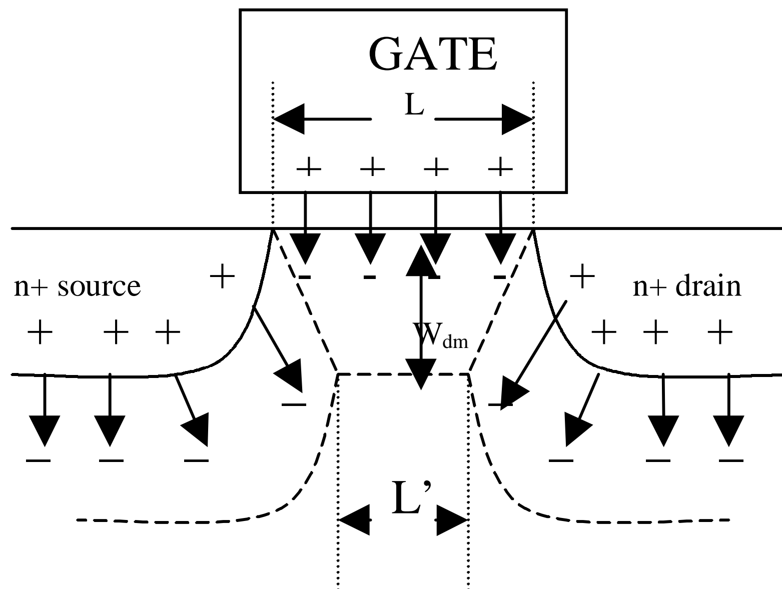

Fig. 13. Schematic diagram for the charge sharing model explaining the reduction of $V_{\rm th}$ source drain depletion region. The bulk charge that needs to be inverted is proportional to the area under the trapezoidal region as given by $Q_{\rm B}{}' \propto W_{\rm dm}(L+L')/2$, which is less than total depletion charge, as in the case of long channel, which is $Q_{\rm B} \propto W_{\rm dm}(L)$.[14]

has to invert less bulk charges resulting in the decrease in the threshold voltage (Fig. 13). In other words, for the same gate voltage, there is more band bending in the Si/SiO_2 interface in a short channel device than in a long channel. Consequently, the threshold voltage is less in a short channel device. The effect of the source-drain depletion region is more in the case of a high drain bias. High drain bias results in more depletion charge in the channel due to drain and source, resulting in a further decrease of threshold voltage. Since, threshold voltage decreases with reduction in channel length, this causes an increase in the subthreshold current.

2.2.5. *Temperature*

The temperature dependency of leakage current is an important consideration, since digital VLSI circuits usually operate at elevated temperatures due to the power dissipation and heat generation of the circuit. $\log(I_{\rm D})$ versus $V_{\rm G}$ shows a linear change in slope S_t with temperature (Fig. 14) as predicted by the logarithm of the subthreshold current model.[6,14] In Fig. 14, S_t varies from 58.2 to 81.9 mV/decade as the temperature increases from $-50°$C to 25°C in a 0.35 μm technology. The increase in $I_{\rm OFF}$ is 0.45 pA to 160 pA for the 20 μm wide device (23 fA/μm to 8 pA/μm). The $I_{\rm OFF}$ increase factor is 356 for this technology. Two parameters increase $I_{\rm OFF}$ as temperature is raised: (1) S_t linearly increases with the Kelvin temperature, and (2) threshold voltage $V_{\rm th}$ decreases. The temperature coefficient of $V_{\rm th}$ was measured at about 0.8 mV/°C for these thin oxides. This allows estimates of $I_{\rm OFF}$ at other temperatures.

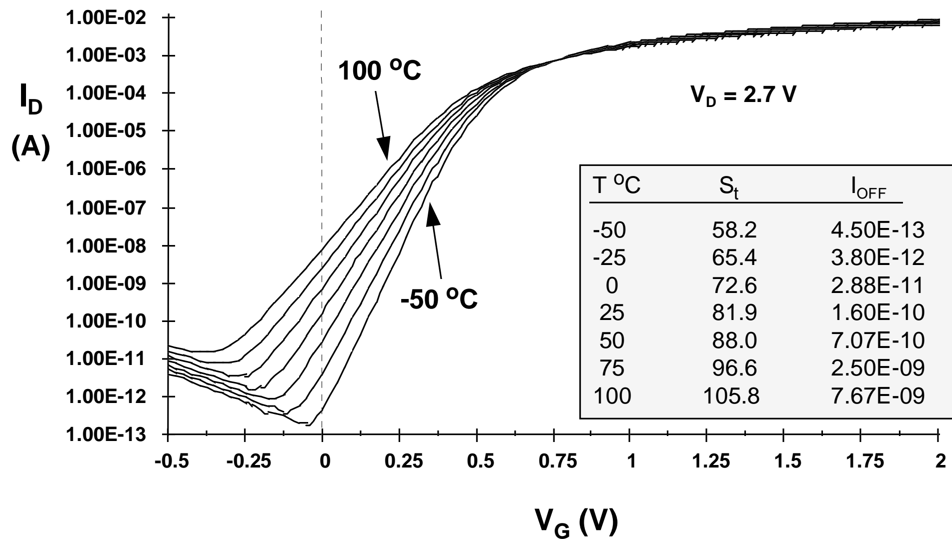

Fig. 14. $I_{\rm D}$ versus $V_{\rm G}$ showing the temperature sensitivity of $I_{\rm OFF}$.[29]

2.3. *Tunneling into and through gate oxide* (I_3)

The reduction of gate oxide thickness results in an increase in the field across the oxide. The high electric field coupled with low oxide thickness results in the tunneling of electron from the substrate to the gate and also from the gate to the substrate through gate oxide resulting in a gate oxide tunneling current.

To understand the phenomenon of tunneling let us consider a MOS capacitor with a heavily doped n^+ type poly-silicon gate and a p-type substrate. Also, for simplicity, let us now focus only on the tunneling of electron. Energy band diagram in flat-band condition is shown in Fig. 15(a), where, Φ_{ox} is the Si–SiO_2 interface barrier height for electron. When a positive bias is applied at gate, the energy band diagram changes as shown in Fig. 15(b). Due to the small oxide thickness, which results in a small width of the potential barrier, the electrons at the strongly inverted surface, can tunnel into or through the SiO_2 layer and hence give rise to the gate current. On the other hand, if a negative gate bias is applied, electron from the n^+ poly-silicon can tunnel into or through the oxide layer and give rise to gate current (Fig. 15(c)).[14]

The mechanism of tunneling between the substrate and the gate poly-silicon can be primarily divided into two parts, namely, (I) Fowler-Nordheim (FN) tunneling and (II) direct tunneling. In the case of Fowler-Nordheim tunneling, electrons tunnel through a triangular potential barrier, whereas, in the case of direct tunneling, electrons tunnel through a trapezoidal potential barrier.

The tunneling probability of an electron depends on

(i) thickness of the barrier,
(ii) barrier height and
(iii) structure of the barrier.

As a result, the tunneling probability of a single electron in FN tunneling and direct tunneling are different resulting in different tunneling current.

2.3.1. *Fowler-Nordheim tunneling*

When the voltage drop across the oxide (V_{ox}), is greater than the barrier height of the electron in a conduction band (ϕ_{ox}) (i.e. $V_{\text{ox}} > \phi_{\text{ox}}$), electrons from the inverted surface tunnel into gate oxide through the conduction band of oxide layer.[14,22] This phenomenon is known as Fowler-Nordheim (FN) tunneling. Figure 16 shows the FN tunneling of electrons from inverted surface to the gate. Since $V_{\text{ox}} > \phi_{\text{ox}}$, electrons has to tunnel through a triangular potential barrier as seen in Fig. 16. Ignoring the effects of finite temperature and image force induced barrier lowering, the current density in FN tunneling is given by[14]

$$J_{\text{FN}} = \frac{q^3 E_{\text{ox}}^2}{16\pi^2 \hbar \phi_{\text{ox}}} \exp\left(-\frac{4\sqrt{2m^*}\phi_{\text{ox}}^{3/2}}{3\hbar q E_{\text{ox}}}\right), \tag{10}$$

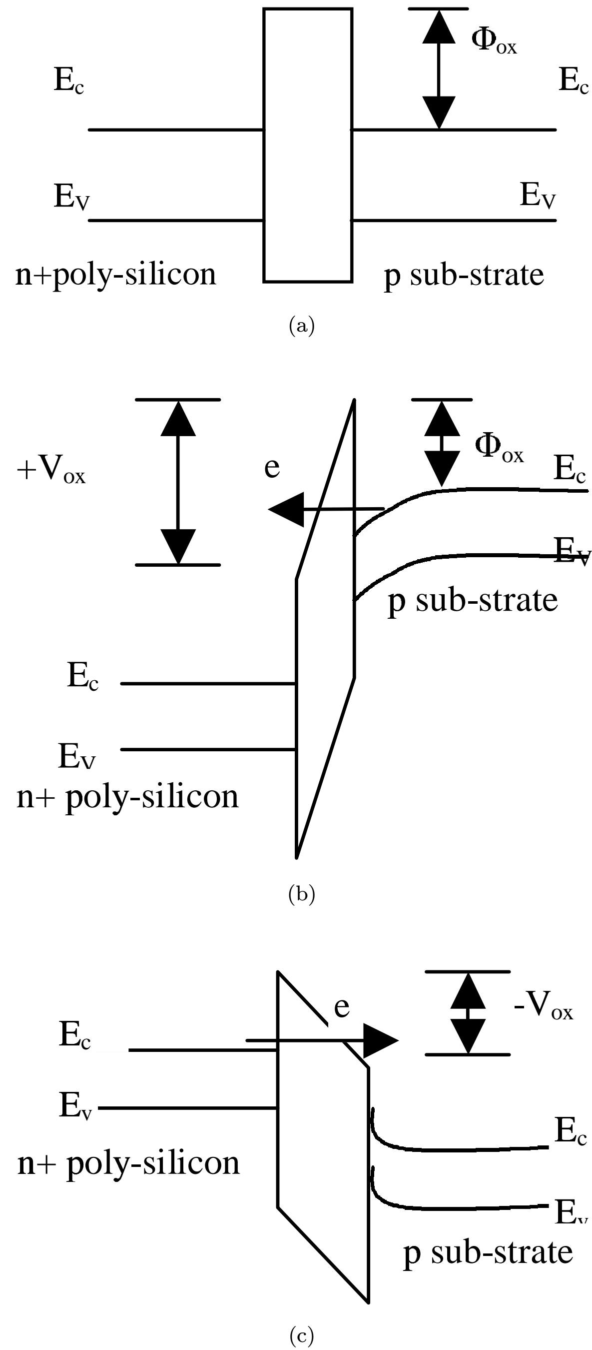

Fig. 15. Tunneling of electron through a MOS capacitor. (a) Energy-band diagram at flat-band condition; (b) energy-band diagram with +ve gate bias showing the tunneling of electron from substrate to gate; and (c) energy-band diagram at −ve gate bias showing tunneling of electron from gate to substrate.[14]

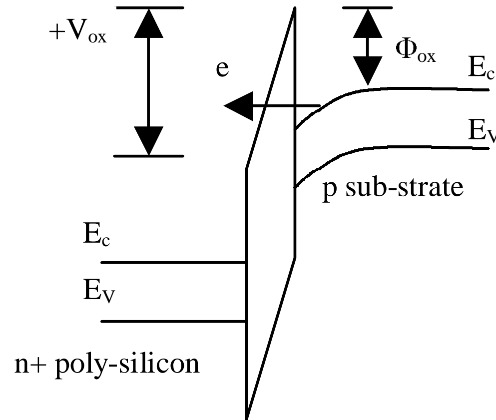

Fig. 16. FN tunneling of electron.

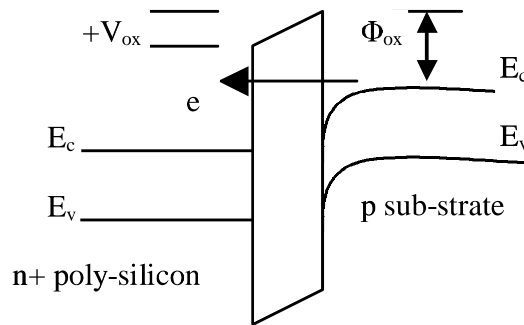

Fig. 17. Direct tunneling of electron.

where E_{ox} is the field across the oxide, ϕ_{ox} is the barrier height for electrons in conduction band and m^* is the effective mass of electrons in conduction band of silicon. FN current equation represents the tunneling through triangular potential barrier and is valid for $V_{\text{ox}} > \phi_{\text{ox}}$.[22] The measured value of FN tunneling current is very small: at oxide field of 8 MV/cm FN tunneling current density is about 5×10^{-7} A/cm^2.[14] Since, $\phi_{\text{ox}} = 3.1$ eV, the short channel device mostly operates with $V_{\text{ox}} < \phi_{\text{ox}}$. Thus, for normal device operation the FN tunneling current is negligible.

2.3.2. *Direct tunneling*

In very thin oxide layer (less than 3–4 nm) electrons from the inverted silicon surface, instead of tunneling into the conduction band of SiO_2, tunnel directly to the gate through the forbidden energy gap of SiO_2 layer.[14] The direct tunneling phenomenon is explained in Fig. 17. Direct tunneling occurs in the region $V_{\text{ox}} <$

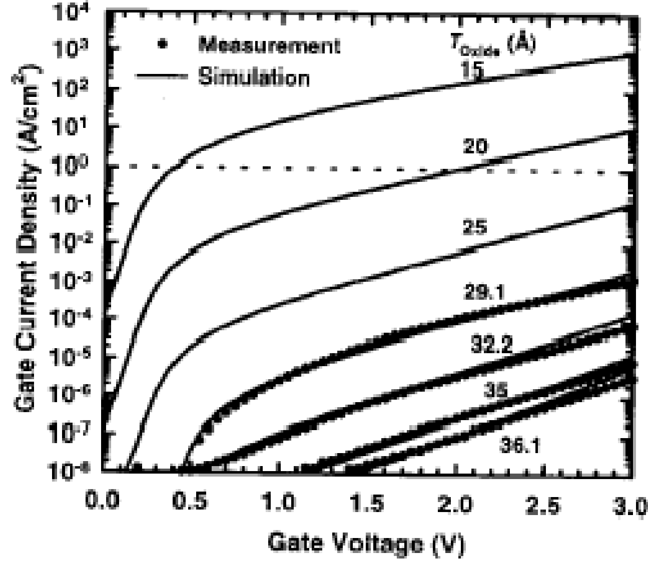

Fig. 18. Measured and simulated direct tunneling current density in thin-oxide poly-silicon gate MOS device.[22]

$\phi_{\rm ox}$.[22] Hence, in the case of direct tunneling, electrons tunnel through a trapezoidal potential barrier instead of a triangular potential barrier as seen in Fig. 17. The equation governing current density in direct tunneling is given by[22]

$$J_{\rm DT} = AE_{\rm ox}^2 \exp\left\{-\frac{B[1-(1-\frac{V_{\rm ox}}{\phi_{\rm ox}})^{3/2}]}{E_{\rm ox}}\right\}, \tag{11}$$

where $A = \frac{q^3}{16\pi^2\hbar\phi_{\rm ox}}$ and $B = \frac{4\sqrt{2m^*}\phi_{\rm ox}^{3/2}}{3\hbar q}$.

The direct tunneling current is significant for low oxide thickness. Figure 18 shows the variation of direct tunneling current with gate voltage.

The potential drop across oxide is obtained from the fact that the applied gate voltage over the flat band voltage drops across the poly-silicon depletion layer, gate oxide and the rest appears as surface potential.

$$V_{\rm gs} = V_{\rm fb} + V_{\rm ox} + \phi_{\rm s} + V_{\rm poly}\,, \tag{12}$$

where

$V_{\rm gs}$ = Applied gate bias,

$\phi_{\rm s}$ = Surface potential,

$V_{\rm poly}$ = Potential drop across the poly-silicon depletion region as given by $\frac{\varepsilon_{\rm ox}^2 E_{\rm ox}^2}{2q\varepsilon_{\rm si}N_{\rm poly}}$ where

$N_{\rm poly}$ = Doping concentration at polysilicon,

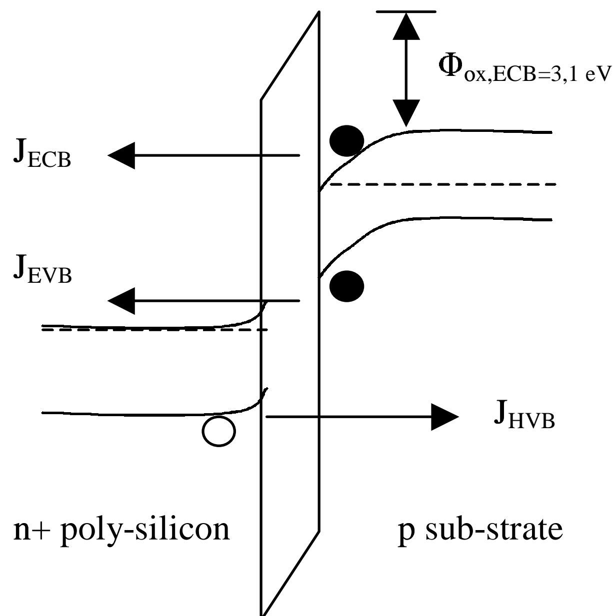

Fig. 19. Three mechanisms for gate leakage.[24,25]

$$\varepsilon_{si} = \text{Permittivity of silicon}\,,$$
$$\varepsilon_{ox} = \text{Permittivity of } SiO_2\,.$$

2.3.2.1. Mechanisms of direct tunneling

There are three major mechanisms for direct tunneling in MOS devices, namely, electron tunneling from conduction band (ECB), electron tunneling from valence band (EVB) and hole tunneling from valance band (HVB)[24,25] (Fig. 19). In NMOS ECB controls the gate-to-channel tunneling current in inversion, whereas gate-to-body tunneling is controlled by EVB in depletion-inversion and ECB in accumulation. In PMOS, HVB controls the gate-to-channel leakage in inversion, whereas, gate-to-body leakage is controlled by the EVB in depletion-inversion and the ECB in accumulation.[24,25] Since, the barrier height for HVB (4.5 eV) is considerably higher as compared to the barrier height for ECB (3.1 eV) and since the effective mass for hole is higher than that of an electron, the tunneling current associated with HVB is much less than the current associated with ECB. This results in lower gate leakage current in PMOS rather than in NMOS.[26]

2.3.2.2. Components of tunneling current

The gate direct tunneling current can be divided into five major components, namely, parasitic leakage current through gate-to-S/D extension overlap region (I_{gso} and I_{gdo}), gate-to-inverted channel current (I_{gc}), part of which goes to source (I_{gcs})

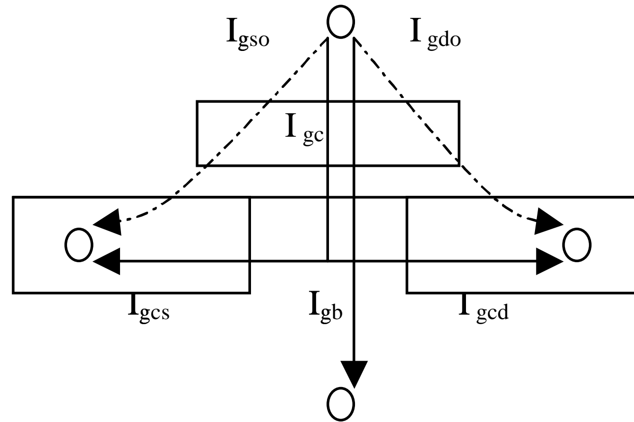

Fig. 20. Components of tunneling current.[24,25]

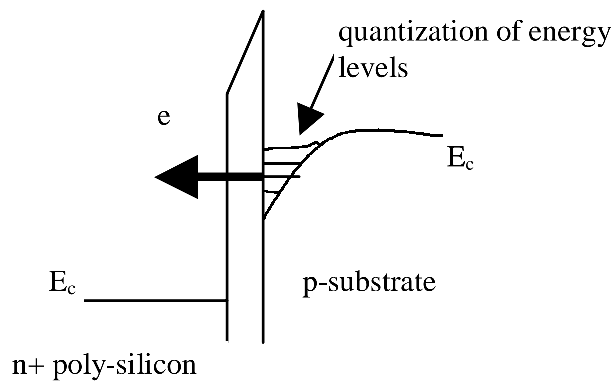

Fig. 21. Quantization of electron energy levels in substrate.

and the rest goes to drain (I_{gcd}), and the gate-to-substrate leakage current (I_{gb}) (Fig. 20).[24,25] The modeling of each of the components can be found in Refs. 24 and 25.

2.3.2.3. Effect of quantization of substrate electron energy

Due to the high substrate doping level and the large electric field at the $\mathrm{Si/SiO_2}$, the quantization of carrier energy occurs within the Si substrate (Fig. 21). This results in less occupied energy states from which the electrons can tunnel. Also, due to the quantization effect, the carrier density in the substrate is different from the classical prediction. With the quantization, the carrier density peaks at a small distance away from the surface and not at the surface as predicted by classical physics. This can be considered as an effective increase in oxide thickness. Thus quantization effect tends to reduce the gate direct tunneling current.[27]

2.3.2.4. Effect of image force induced barrier lowering

The emission of electron from Si to SiO_2 causes a build up of image charge at the oxide side Si/SiO_2 interface which results in a reduction in the barrier height at the Si/SiO_2 interface from $\phi_{ox} = 3.1$ eV by an amount of $\Delta\phi$, where

$$\Delta\phi = \sqrt{\frac{q^3 E_{ox}}{4\pi\varepsilon_{ox}}}, \quad (\varepsilon_{ox} = \text{permittivity of } SiO_2)\,. \tag{13}$$

This is called the image-force-induced-barrier lowering effect.[14] Since, it modulates the barrier height, it also modulates the gate tunneling current, since tunneling depends exponentially on ϕ_{ox}.

2.4. *Injection of hot carriers from substrate to gate oxide* (I_4)

In a short channel transistor, due to the high electric field near the Si/SiO_2 interface, electrons or holes can gain sufficient energy from the electric field to cross the interface potential barrier and enter into the oxide layer (Fig. 22). This effect is known as the hot carrier injection. The injection from Si to SiO_2 is more likely for an electron than for a hole as the electron has a lower effective mass than that of a hole. In addition, the barrier height for hole is more (4.5 eV) than the barrier height for an electron (3.1 eV).[14]

2.5. *Gate induced drain leakage* (I_5)

Gate induced drain leakage (GIDL) is due to the high field effect in the drain junction of a MOS transistor. When the gate is biased to cause an accumulation layer to

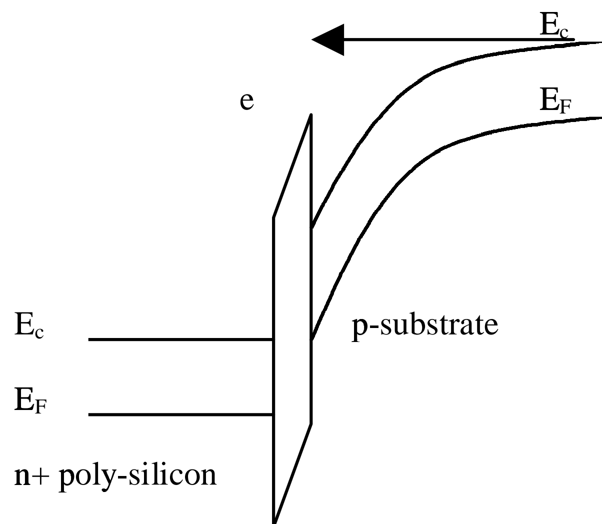

Fig. 22. Injection of hot electron from substrate to oxide.

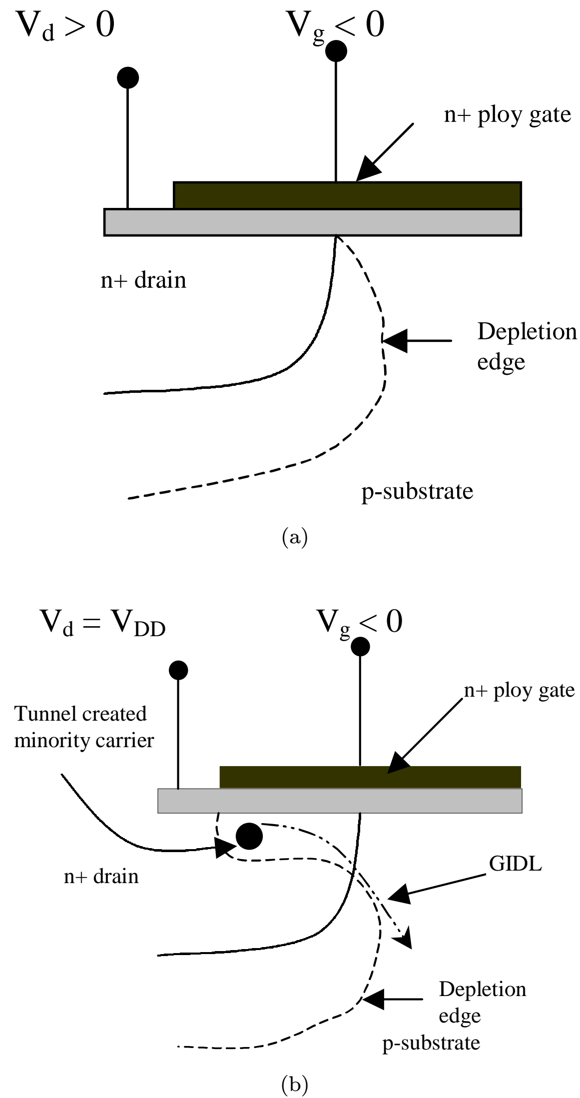

Fig. 23. Condition of the depletion region near the drain-gate overlap region of MOS transistor when (a) surface is accumulated with low negative gate bias, (b) n^+ region is depleted or inverted with high negative gate bias.

form at the silicon surface, the silicon surface under the gate has almost the same potential as the *p*-type substrate. Due to the presence of accumulated holes at the surface, the surface behaves like a more heavily doped *p*-region than the substrate. Thus the depletion layer at the surface is much narrower than elsewhere (Fig. 23(a)). The narrowing of depletion layer at or near surfaces causes field crowding or increases in the local electric field, thereby enhancing the high field effects near that region.[14] When the negative gate bias is large (i.e. the gate at zero or negative and

the drain at V_{DD}), the n^+ drain region under the gate can be depleted and even inverted as shown in Fig. 23(b). This causes more field crowding and increases in the peak field, resulting in dramatic increases of high field effects like avalanche multiplication, and band-to-band tunneling.[14] The possibility of tunneling via near surface trap also increases. As a result of these effects, minority carriers are emitted in the drain region underneath the gate. Since the substrate is at a lower potential for minority carriers, the minority carriers that have been accumulated or formed at the drain depletion region underneath the gate are swept laterally to the substrate, completing a path for the gate induced drain leakage (GIDL).[20] Thinner oxide thickness and higher V_{DD} (higher potential between gate and drain) enhance the electric field, hence increases the GIDL. The impact of drain (and well) doping on GIDL is rather complicated. At low drain doping values, the electric field is not high enough to cause tunneling. For very high drain doping, the depletion width and tunneling will be limited, causing less GIDL. Hence, GIDL is worse for drain doping values in between the above extremes. Very high and abrupt drain doping is preferred for minimizing GIDL as it provides lower series resistance required for the high transistor drive current.[16]

2.6. *Punch-through* (I_6)

In short channel devices, due to the proximity of drain and source, the depletion regions at the drain-substrate and substrate-source junction extend into the channel. As the channel length is reduced, if the doping is kept constant, the separation between the depletion region boundaries decreases. The increase in the reverse bias across the junctions (with increase in V_{ds}) also leads to the boundaries being pushed further away from the junction and nearer to each other. When the combination of channel length and reverse bias causes the depletion regions to merge, then punch-through is said to have occurred. In sub-micron MOSFETs a V_{th}-adjust implant is used to cause higher doping at the surface rather than in the bulk. This causes greater expansion of the depletion region below the surface (due to the smaller doping there) than at the surface. Thus punch-through occurs below the surface.[20] An increase in the drain voltage beyond the value required to establish the punch-through lowers the potential barrier for the majority carriers in the source. Thus, more of these carriers cross the energy barriers and enter into the substrate. The drain collects some of them. The net effect is an increase in the subthreshold current. Furthermore, the punch-through causes a decrease in the subthreshold slope. The device parameter commonly used to characterize punch-through is the punch-through voltage V_{PT}, which estimates the value of V_{ds} for which punch-through occurs (i.e. subthreshold current reaches a particular value) with $V_{\mathrm{gs}} = 0$. It is roughly estimated as the value of the V_{ds} for which the sum of the width of drain and source depletion region is equal to effective channel length

$$V_{\mathrm{PT}} \propto N_{\mathrm{B}}(L - W_j)^3 \,, \tag{14}$$

where

$$N_B = \text{Doping concentration at the bulk},$$
$$L = \text{Channel length},$$
$$W_j = \text{Junction width}.$$

The most suitable method for controlling punch-through is to use additional implants. A layer of higher doping at a depth equal to that of the bottom of the junction depletion regions is one possible solution. Another approach could be to form a halo at the leading edge of drain-and-source junction.[20]

3. Conclusion

With the continuous scaling of CMOS devices, leakage current is becoming a major contributor to the total power consumption. In current deep submicron devices with low threshold voltages, subthreshold leakage has become the dominant source of leakage and is expected to increase with technology scaling. Gate oxide tunneling is likely to become a problem in the future as the oxide thickness continues to shrink. Gate induced drain leakage may also become a concern. To manage the increasing leakage in future CMOS technologies, solutions for leakage reduction have to be sought both at the circuit and process technology levels.

Acknowledgment

This work was supported in part by Semiconductor Research Corporation, DARPA, Intel, and IBM.

References

1. V. De and S. Borkar, "Technology and design challenges for low power and high performance", *Proc. Int. Symp. Low Power Electron. Design*, August 1999, pp. 163–168.
2. A. Righter, J. Soden, and R. Beegle, "High resolution I_{DDQ} characterization and testing — practical issues", *Proc. Int. Test Conf.*, October 1996, pp. 259–268.
3. C. Mead, "Scaling of MOS technology to submicrometer feature sizes", *Analog Integrated Circuit Signal Process* **6** (1994) 9–25.
4. R. Dennard *et al.*, "Design of ion-implanted MOSFET's with very small physical dimensions", *IEEE J. Solid State Circuits*, October 1974, pp. 256.
5. Y. Tsividis, *Operation and Modeling of the MOS Transistor*, McGraw-Hill, New York, 1987.
6. R. Pierret, *Semiconductor Device Fundamentals*, Addison-Wesley, Reading, MA, 1996.
7. J. Brews, *High Speed Semiconductor Devices*, ed. S. M. Sze, New York, USA: John Wiley & Sons, 1990, chapter 3.
8. 2001 International Technology Roadmap for Semiconductors, http://public.itrs.net/.
9. S. Thompson, P. Packan, and M. Bohr, "Linear versus saturated drive current: tradeoffs in super steep retrograde well engineering", *Symp. VLSI Technol.*, 1996, pp. 154–155.

10. S. Venkatesan, J. W. Lutze, C. Lage, and W. J. Taylor, "Device drive current degradation observed with retrograde channel profiles", *Int. Electron Devices Meeting*, 1995, pp. 419–422.
11. J. Jacobs and D. Antoniadis, "Channel profile engineering for MOSFET's with 100 nm channel lengths", *IEEE Trans. Electron Devices* **42** (1995) 870–875.
12. M. Cao, P. Griffin, P. V. Voorde, C. Diaz, and W. Greene, "Transient-enhanced diffusion of iridium and its effects on electrical characteristics of deep sub-micron nMOSFETs", *Digest Tech. Papers Symp. VLSI Technol.*, 1997, pp. 85–86.
13. A. S. Grove, *Physics and Technology of Semiconductor Devices*, John Wiley & Sons, New York, USA 1967.
14. Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, New York, 1998.
15. R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFETS with very small physical dimensions", *IEEE J. Solid-State Circuits* **SC-9** (1974) 256.
16. V. De, Y. Ye, A. Keshavarzi, S. Narendra, J. Kao, D. Somasekhar, R. Nair, and S. Borkar, "Techniques for leakage power reduction", *Design of High-Performance Microprocessor Circuits*, eds. A. Chnadrakasan, W. J. Bowhill and F. Fox, IEEE Press, Piscataway, NJ, USA, 2001, chapter 3, pp. 46–62.
17. S. Chung and C.-T. Li, "An analytical threshold-voltage model of trench-isolated MOS devices with nonuniformly doped substrates", *IEEE Trans. Electron Devices* **39** (1992) 614–622.
18. D. Fotty, *MOSFET Modelling with SPICE*, Prentice Hall PTR, New Jersey, USA.
19. BSIM3v3.2.2 MOSFET Model BSIM Group, University of California Berkeley. http://www-device.eecs.berkeley.edu/∼bsim3/.
20. K. Roy and S. C. Prasad, *Low-Power CMOS VLSI Circuit Design*, Wiley Interscience Publications, New York, USA, 2000.
21. J. Mandelman and J. Alsmeir, "Anomalous narrow channel effect in trench-isolated burried channel p-MOSFETS", *IEEE Electron Device Lett.* **15** (1994) 496–498.
22. K. Schuegraf and C. Hu, "Hole injection SiO_2 breakdown model for very low voltage lifetime extrapolation", *IEEE Trans. Electron Device* **41** (1994) 761–767.
23. S. Lo *et al.*, Modeling and characterization of n^+- and p^+-polysilicon-gated ultra thin oxides (21–26 A^0)", *Symp. VLSI Technol.*, 1997, pp. 149–150.
24. BSIM4.2.1 MOSFET Model, BSIM Group, University of California Berkeley. http://www-device.eecs.berkeley.edu/∼bsim3/
25. K. Cao, W.-C. Lee, W. Liu, X. Jin, P. Su, S. Fung, J. An, B. Yu, and C. Hu, "BSIM4 gate leakage model including source drain partiotion", *Tech. Digest Int. Electron Devices Meeting*, 2000, pp. 815–818.
26. F. Hamzaoglu and M. Stan, "Circuit-level techniques to control gate leakage for sub-100 nm CMOS", *Int. Symp. Low Power Design*, 2002.
27. N. Yang, W. Henson, and J. Hauser, "Modeling study of ultrathin gate oxides using tunneling current and capacitance-voltage measurement in MOS devices", *IEEE Trans. Electron Devices* **46** (1999) 1464–1471.
28. Y. Taur, "CMOS scaling and issues in sub-0.25 μm systems", *Design of High-Performance Microprocessor Circuits*, eds. A. Chnadrakasan, W. J. Bowhill and F. Fox, IEEE Press, Piscataway, NJ, USA, 2001, chapter 2, pp. 27–45.
29. A. Keshavarzi, K. Roy, and C. F. Hawkins, "Intrinsic leakage in low power deep submicron CMOS ICs", *Int. Test Conf.*, 1997, pp. 146–155.