Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits

KAUSHIK ROY, FELLOW, IEEE, SAIBAL MUKHOPADHYAY, STUDENT MEMBER, IEEE, AND
HAMID MAHMOODI-MEIMAND, STUDENT MEMBER, IEEE

Contributed Paper

High leakage current in deep-submicrometer regimes is becoming a significant contributor to power dissipation of CMOS circuits as threshold voltage, channel length, and gate oxide thickness are reduced. Consequently, the identification and modeling of different leakage components is very important for estimation and reduction of leakage power, especially for low-power applications. This paper reviews various transistor intrinsic leakage mechanisms, including weak inversion, drain-induced barrier lowering, gate-induced drain leakage, and gate oxide tunneling. Channel engineering techniques including retrograde well and halo doping are explained as means to manage short-channel effects for continuous scaling of CMOS devices. Finally, the paper explores different circuit techniques to reduce the leakage power consumption.

Keywords**—***Channel engineering, CMOS, dynamic $V_{\rm dd}$, dynamic $V_{\rm th}$, gate leakage, leakage current, low-leakage memory, multiple $V_{\rm dd}$, multiple $V_{\rm th}$, scaling, stacking effect, subthreshold current, tunneling.*

I. Introduction

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

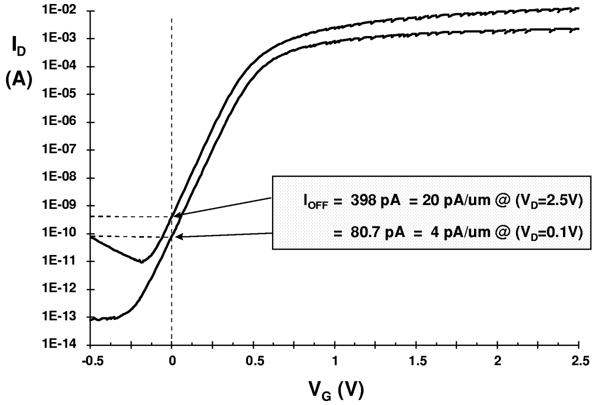


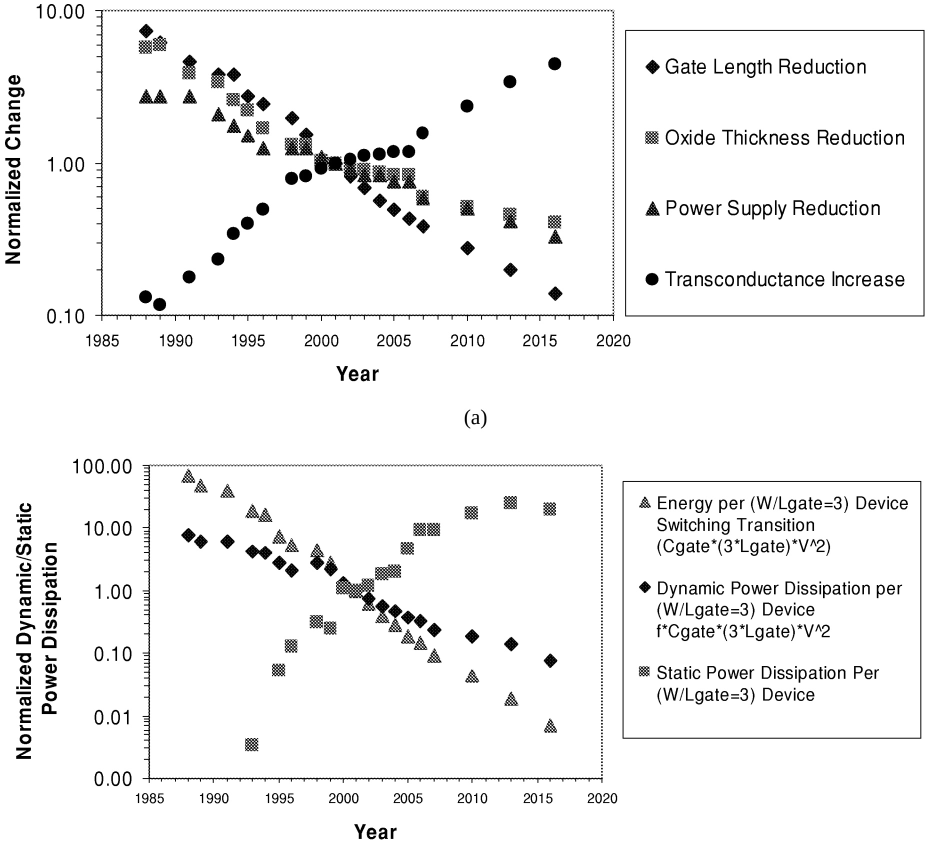
Fig. 1 Log (I_D) versus V_G at two different drain voltages for 20 × 0.4-μm n-channel transistor in a 0.35-μm CMOS process [2].

Fig. 1 shows a typical curve of drain current (I_D) versus gate voltage (VG) in logarithmic scale [2]. It allows measurement of many device parameters such as $I_{\rm OFF}$, $V_{\rm th}$, and subthreshold slope (S_t), that is, the slope of V_G versus I_D in the weak inversion state. Transistor off-state current $(I_{\rm OFF})$ is the drain current when the gate voltage is zero. The n-channel transistor in Fig. 1 has an $I_{\rm OFF}$ of 20 and 4 pA$/\mu$m at the drain voltage of 2.5 and 0.1 V, respectively. $I_{\rm OFF}$ 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 leakage from the drain-well and well-substrate reverse-bias pn junctions [2]. Short-channel transistors require lower power supply levels to reduce their internal electric fields and power consumption. This forces a reduction in the threshold voltage that causes a substantially large increase in $I_{\rm OFF}$ [1]. This increase is due to the weak inversion state

Manuscript received July 10, 2002; revised November 5, 2002. This work was supported in part by Semiconductor Research Corporation, in part by Defense Advanced Research Projects Agency, Intel, and IBM.

The authors are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907-1285 USA (e-mail: kaushik@ecn.purdue.edu; sm@ecn.purdue.edu; mahmoodi@ecn.purdue.edu).

Digital Object Identifier 10.1109/JPROC.2002.808156



(a)

(b)

Fig. 2 ITRS projections for transistor scaling trends and power consumption: (a) physical dimensions and supply voltage and (b) device power consumption [6].

leakage and is a function of V_{th}. In this paper, we explore all leakage mechanisms contributing to the off-state current (not just the current from the drain terminal). Other leakage mechanisms are peculiar to the small geometries themselves. As the drain voltage increases, the drain to channel depletion region widens, resulting in a significant increase in the drain current. This increase in I_{OFF} is typically due to channel surface current caused by drain-induced barrier lowering (DIBL) or due to deep channel punchthrough currents [3]–[5]. Moreover, as the channel width decreases, the threshold voltage and the off current both get modulated by the width of the transistor, giving rise to significant narrow-width effect. All these adverse effects which cause threshold voltage reduction (leakage current increase) in scaled devices are called short-channel effects (SCE). To maintain a reasonable SCE immunity while scaling down the channel length, oxide thickness has to be reduced nearly in proportion to the channel length. Decrease in oxide thickness results in increase in the 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 gate oxide tunneling and injection of hot carrier from substrate to the gate oxide. Gate-induced drain leakage (GIDL) is another significant leakage mechanism, resulting due to the depletion at the drain surface below the gate-drain overlap region. Fig. 2 shows projections for transistor physical dimensions, supply voltage, and device power consumption according to the International Technology Roadmap for Semiconductors (ITRS) [6]. All the parameters are normalized to their values in the year 2001. As shown in Fig. 2(b), due to the substantial increase in the leakage current, the static power consumption is expected to exceed the switching component of the power consumption unless effective measures are taken to reduce the leakage power.

Due to adverse SCEs, the channel length cannot be arbitrarily reduced even if allowed by lithography. For digital applications, the most undesirable SCE 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 SCEs under control at very small dimensions. In addition to 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. Supersteep retrograde wells and halo implants have been used as a means to scale

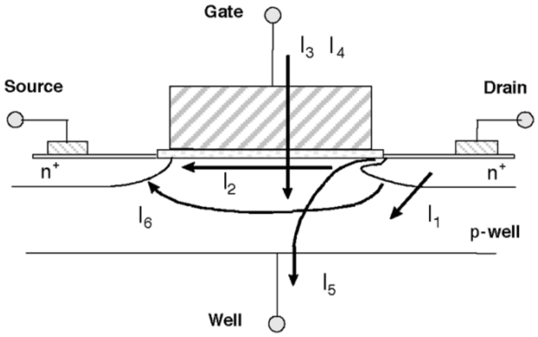


Fig. 3 Summary of leakage current mechanisms of deep-submicrometer transistors.

the channel length and increase the transistor drive current without causing an increase in the off-state leakage current [7]–[10].

This paper is organized as follows. In Section II, different leakage current components and mechanisms in deep-submicrometer transistors are explained, which is essential to guide solutions for reducing power and leakage per transistor. Device options for leakage reduction, which are based on channel engineering, are explained in the first part of Section III. The second part of Section III explores different circuit techniques for leakage control in logic and memory. Finally, the conclusion of the paper appears in Section IV.

II. Transistor Leakage Mechanisms

We describe six short-channel leakage mechanisms as illustrated in Fig. 3. I_1 is the reverse-bias pn junction leakage; I_2 is the subthreshold leakage; I_3 is the oxide tunneling current; I_4 is the gate current due to hot-carrier injection; I_5 is the GIDL; and I_6 is the channel punchthrough current. Currents I_2, I_5, and I_6 are off-state leakage mechanisms, while I_1 and I_3 occur 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.

A. pn Junction Reverse-Bias Current (I_1)

Drain and source to well junctions are typically reverse biased, causing pn junction leakage current. A reverse-bias pn junction leakage (I_1) has two main components: one is minority carrier diffusion/drift near the edge of the depletion region; the other is due to electron-hole pair generation in the depletion region of the reverse-biased junction [12]. For an MOS transistor, additional leakage can occur between the drain and well junction from gated diode device action (overlap of the gate to the drain-well pn junctions) or carrier generation in drain to well depletion regions with influence of the gate on these current components [13]. pn junction reverse-bias leakage $(I_{\rm REV})$ is a function of junction area and doping concentration [12]. If both n and p regions are heavily doped (this is the case for advanced MOSFETs using heavily doped shallow junctions and halo doping for better SCE), band-to-band tunneling (BTBT) dominates the pn junction leakage [14]. This leakage mechanism is explained in Section II-A1.

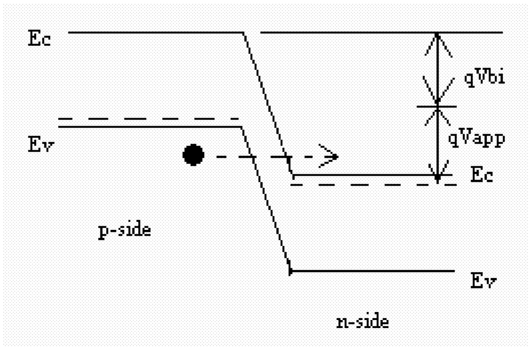


Fig. 4 BTBT in reverse-biased pn junction [14].

1) Band-to-Band Tunneling Current: High electric field $(>10^6$ V/cm$)$ across the reverse-biased pn junction causes significant current to flow through the junction due to tunneling of electrons from the valence band of the p region to the conduction band of the n region, as shown in Fig. 4 [14]. From Fig. 4, it is evident that for the tunneling to occur, the total voltage drop across the junction has to be more than the band gap. The BTBT current in silicon involves the emission or absorption of phonons, since silicon is an indirect band gap semiconductor. The tunneling current density is given by [14]

$$J_{b-b} = A\frac{EV_{\rm app}}{E_g^{1/2}}\exp\left(-B\frac{E_g^{3/2}}{E}\right)$$
$$A = \frac{\sqrt{2m^*}q^3}{4\pi^3\hbar^2}, \text{ and } B = \frac{4\sqrt{2m^*}}{3q\hbar} \tag{1}$$

where m^* is effective mass of electron; E_g is the energy-band gap; $V_{\rm app}$ is the applied reverse bias; E is the electric field at the junction; q is the electronic charge; and $\hbar$ is $1/2\pi$ times Planck's constant.Assuming a step junction, the electric field at the junction is given by [14]

$$E = \sqrt{\frac{2qN_aN_d(V_{\rm app}+V_{bi})}{\varepsilon_{\rm si}(N_a+N_d)}} \tag{2}$$

where N_a and N_d are the doping in the p and n side, respectively; $\epsilon_{\rm si}$ is permittivity of silicon; and V_{bi} is the built in voltage across the junction. In scaled devices, high doping concentrations and abrupt doping profiles cause significant BTBT current through the drain-well junction.

B. Subthreshold Leakage (I_2)

Subthreshold or weak inversion conduction current between source and drain in an MOS transistor occurs when gate voltage is below $V_{\rm th}$ [15]. The weak inversion region is seen in Fig. 1 as the linear region of the curve (semilog plot). In the weak inversion, the minority carrier concentration is small, but not zero. Fig. 5 shows the variation of minority carrier concentration along the length of the channel for an n-channel MOSFET biased in the weak inversion region. 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. For such weak inversion condition, $V_{\rm ds}$ drops almost entirely across the reverse-biased substrate-drain pn junction.

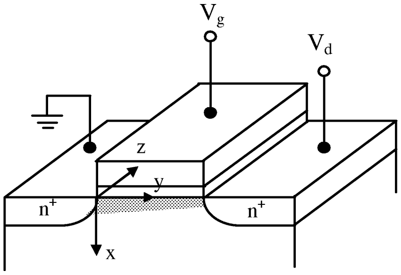


Fig. 5 Variation of minority carrier concentration in the channel of a MOSFET biased in the weak inversion.

As a result, the variation of the electrostatic potential ϕ_s at the semiconductor surface along the channel (the y axis) is small. The y component of the electric field vector $\mathbf{E}(\mathbf{E}_y)$, 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, the 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 driving voltage on the gate and the drain current is a straight line in a semilog plot of I_D versus V_g (see Fig. 6). Weak inversion typically dominates modern device off-state leakage due to the low $V_{\rm th}$. The weak inversion current can be expressed based on the following [15]:

$$I_{\rm ds} = \mu_0 C_{\rm ox}\frac{W}{L}(m-1)(v_T)^2 \times e^{(V_g - V_{\rm th})/mv_T} \times \left(1 - e^{-v_{\rm DS}/v_T}\right) \quad (3)$$

where

$$m = 1 + \frac{C_{\rm dm}}{C_{\rm ox}} = 1 + \frac{\frac{\varepsilon_{\rm si}}{W_{\rm dm}}}{\frac{\varepsilon_{\rm ox}}{t_{\rm ox}}} = 1 + \frac{3t_{\rm ox}}{W_{\rm dm}} \quad (4)$$

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

In long-channel devices, the subthreshold current is independent of the drain voltage for $V_{\rm DS}$ larger than a few v_T. On the other hand, the dependence on the gate voltage is exponential, as illustrated in Fig. 6 [16]. The inverse of the slope of the $\log_{10}(I_{\rm ds})$ versus $V_{\rm gs}$ characteristic is called the subthreshold slope (S_t) [15] and is given by

$$S_t = \left(\frac{d\left(\log_{10} I_{\rm ds}\right)}{dV_{\rm gs}}\right)^{-1} = 2.3\frac{mkT}{q} = 2.3\frac{kT}{q}\left(1 + \frac{C_{\rm dm}}{C_{\rm ox}}\right). \quad (5)$$

Subthreshold slope indicates how effectively the transistor can be turned off (rate of decrease of $I_{\rm OFF}$) when $V_{\rm gs}$ is decreased below $V_{\rm th}$. As device dimensions and the supply voltage are scaled down to enhance performance, power efficiency, and reliability, subthreshold characteristics may limit the scalability of the supply voltage. The parameter S_t is measured in millivolts per decade of the drain current. For the limiting case of $t_{\rm ox} \to 0$ and at room temperature, $S_t \approx 60$ mV/decade. Typical S_t values for a bulk CMOS process can range from 70 to 120 mV/decade. A low value for subthreshold slope is desirable. It can be noted from the preceding expression that S_t can be made smaller by using a thinner oxide (insulator) layer to reduce $t_{\rm ox}$ or a lower substrate doping concentration (resulting in larger $W_{\rm dm}$). Changes in operating conditions—namely, lower temperature or a substrate bias—also modifies S_t.

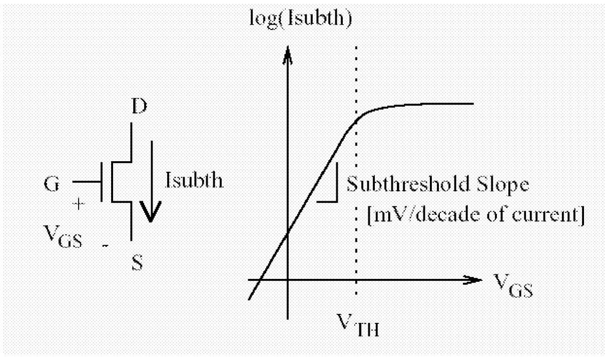
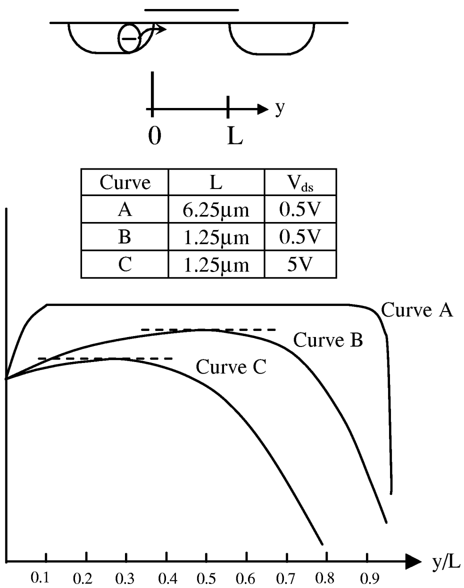


Fig. 6 Subthreshold leakage in a negative-channel metal–oxide–semiconductor (NMOS) transistor.

1) Drain-Induced Barrier Lowering: In long-channel devices, the source and drain are separated far enough that their depletion regions have no effect on the potential or field pattern in most part of the device. Hence, for such devices, 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 have 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 DIBL. One way to describe it is to consider the energy barrier at the surface between the source and drain, as shown in Fig. 7 [17]. Under off conditions, this potential 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 $V_{\rm ds}$. However, the barrier of a short-channel device reduces with an increase in the drain voltage, which in turn increases the subthreshold current due to lower threshold voltage.

DIBL occurs when the depletion regions of the drain and the source interact with each other near the channel surface to lower the source potential barrier. When a high drain voltage is applied to a short-channel device, it lowers the barrier height, resulting in further decrease of the threshold voltage. The source then injects carriers into the channel surface (independent of gate voltage). DIBL is enhanced at high drain voltages and shorter channel lengths. The surface DIBL typically occurs before the deep bulk punchthrough. Ideally,



Curve	L	V_{ds}
A	6.25μm	0.5V
B	1.25μm	0.5V
C	1.25μm	5V

Fig. 7 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 same for all three cases [17].

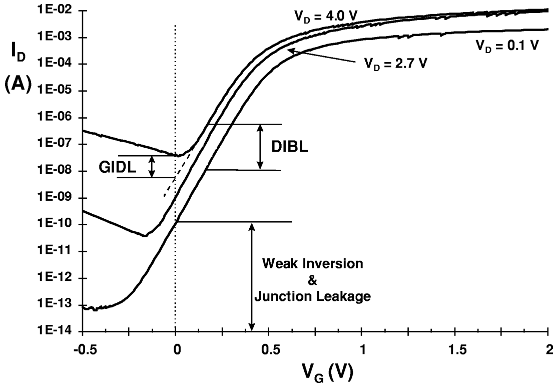


Fig. 8 n channel I_D vs. V_G showing DIBL, GIDL, weak inversion, and pn junction reverse-bias leakage components [11].

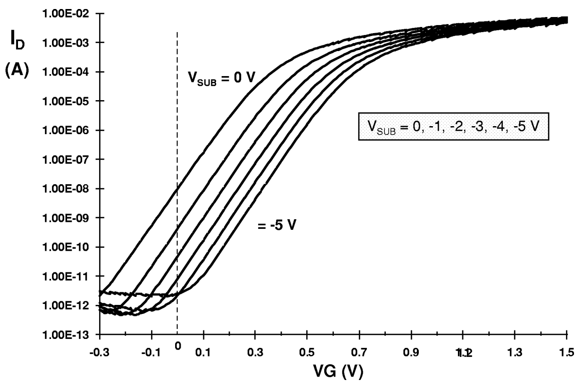


Fig. 9 n channel $\log(I_D)$ versus V_G for six substrate biases on a 0.35-μm logic process technology ($V_D = 2.7$ V) [11].

DIBL does not change the subthreshold slope (S_t), but does lower $V_{\rm th}$. Higher surface and channel doping and shallow source/drain junction depths reduce the DIBL effect on the subthrshold leakage current [17], [18]. Fig. 8 illustrates the DIBL effect as it moves the $I_D - V_G$ curve up and to the left as the drain voltage increases. DIBL can be measured at constant V_G as the change in I_D for a change in V_D [11].

2) Body Effect: Reverse biasing well-to-source junction of a MOSFET transistor widens the bulk depletion region and increases the threshold voltage [19]. The effect of body bias can be considered in the threshold voltage equation [20]

$$V_{\rm th} = V_{\rm fb} + 2\psi_B + \frac{\sqrt{2\varepsilon_{\rm si} q N_a \left(2\psi_B + V_{\rm bs}\right)}}{C_{\rm ox}} \qquad (6)$$

where $V_{\rm fb}$ is the flat-band voltage; N_a is the doping density in the substrate; and $\psi_B = (KT/q)\ln(N_a/n_i)$ is the difference between the Fermi potential and the intrinsic potential in the substrate. The slope of $V_{\rm th}$ versus $V_{\rm bs}$ curve is therefore

$$\frac{dV_{\rm th}}{dV_{\rm bs}} = \frac{\sqrt{\frac{\varepsilon_{\rm si} q N_a}{2(2\psi_B + V_{\rm bs})}}}{C_{\rm ox}} \qquad (7)$$

which is referred to as the substrate sensitivity. It can be seen from (7) that the substrate sensitivity is higher for higher bulk doping concentration, and the substrate sensitivity decreases as the substrate reverse bias increases. At $V_{\rm bs} = 0$, the substrate sensitivity is $C_{\rm dm}/C_{\rm ox}$ or $m - 1$ (4). Therefore, m is also called body effect coefficient.

Fig. 9 shows suppression in 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) [11]. Virtually no change is seen in the subthreshold slope S_t at different substrate biases. An important observation from Fig. 9 is that as $V_{\rm th}$ increases, because of applied reverse substrate bias and a shift in the I–V curve, $I_{\rm OFF}$ decreases.

The subthreshold leakage of an MOS device including weak inversion, DIBL, and body effect, can be modeled as [21]

$$I_{\rm subth} = A \times e^{1/mv_T\left(V_G - V_S - V_{\rm th0} - \gamma' \times V_S + \eta V_{\rm DS}\right)} \times \left(1 - e^{-v_{\rm DS}/v_T}\right) \qquad (8)$$

where

$$A = \mu_0 C'_{\rm ox} \frac{W}{L_{\rm eff}} (v_T)^2 e^{1.8} e^{-\Delta V_{\rm th}/\eta v_T} \qquad (9)$$

$V_{\rm th0}$ is the zero bias threshold voltage, and $v_T = KT/q$ is the thermal voltage. The body effect for small values of source to bulk voltages is linear and is represented by the term $\gamma' V_S$ in (7), 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 of the transistor. $\Delta V_{\rm TH}$ is a term introduced to account for transistor-to-transistor leakage variations.

3) Narrow-Width Effect: The decrease in gate width modulates the threshold voltage of a transistor, and thereby modulates the subthreshold leakage. There are mainly three ways that narrow width modulates the threshold voltage.

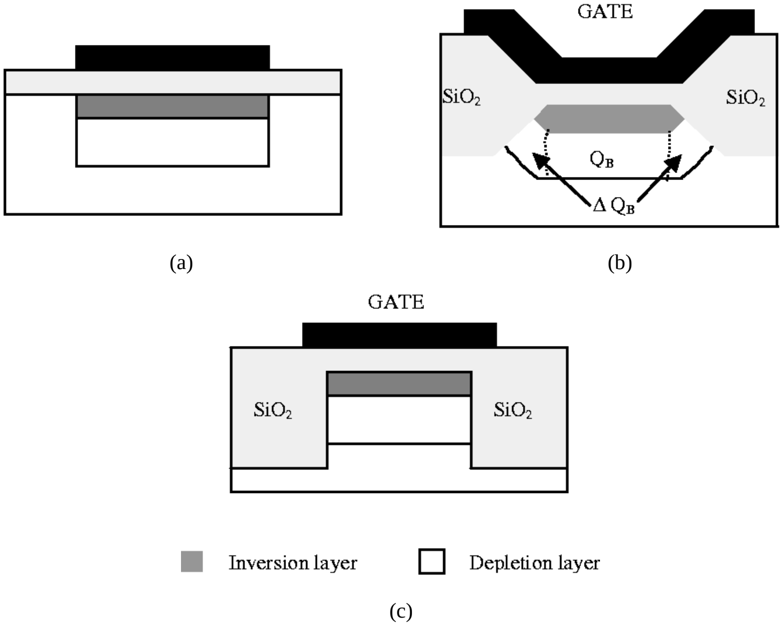


Fig. 10 Three types of device structures and associated inversion–depletion layer. (a) Large-geometry MOSFET. (b) LOCOS gate MOSFET. (c) Trench isolated MOSFET [22].

First, let us consider the local oxide isolation (LOCOS) gate MOSFET. In the LOCOS gate MOSFET, the existence of the fringing field causes the gate-induced depletion region to spread outside the defined channel width and under the isolations as shown in Fig. 10(b). This results in an increase of the total depletion charge in the bulk region above its expected value. The threshold voltage of MOS can be defined using depletion approximation as [22]

$$V_{\mathrm{th}} = V_{\mathrm{fb}} + \phi_s + \frac{Q_B}{C_{\mathrm{ox}}} \tag{10}$$

where V_{fb} is the flat-band voltage; ϕ_s is the surface potential; C_{ox} is the capacitance across the oxide; and Q_B is the depletion charge in the bulk. Due to narrow-width effect, Q_B increases by ΔQ_B as shown in Fig. 10(b). This effect becomes more substantial as the channel width decreases, and the depletion region underneath the fringing field is comparable to the classical depletion formed by the vertical field. This results in increase of threshold voltage due to narrow-channel effect [23], [24]. This narrow-width effect can be modeled as an increase in V_{th} by the amount V_{NCE} given by [25]

$$V_{\mathrm{NCE}} = \frac{\pi q N_{\mathrm{sub}} x_{d,\max}^2}{2 C_{\mathrm{ox}} W_{\mathrm{eff}}} = 3\pi \frac{t_{\mathrm{ox}}}{W_{\mathrm{eff}}} \phi_s \tag{11}$$

where N_{sub} is the substrate doping; $x_{d,\max}$ is the maximum vertical depletion width; C_{ox} is the capacitance across the oxide; W_{eff} is the effective width; t_{ox} is the oxide thickness; and ϕ_s is the surface potential. A more accurate model can be found in [24].

The second way that narrow-width modulates the threshold voltage is due to the fact that the channel doping is higher along the width dimension in LOCOS gates. Due to the channel stop, dopants encroach under the gate. Hence, a higher voltage is needed to completely invert the channel [26].

A more complex effect is seen in trench isolation devices, known as inverse-narrow-width effect. In the case of trench isolation devices, depletion layer cannot spread under the oxide isolation [see Fig. 10(c)]. Hence, the total depletion charge in the bulk does not increase $(\Delta Q_B \approx 0)$, thereby eliminating the increase in the threshold voltage. On the other hand, due to the two-dimensional (2-D) field-induced edge-fringing effect at the gate edge, formation of an inversion layer at the edges occurs at a lower voltage than the voltage required at the center. Moreover, the overall gate capacitance (C_T) now includes the sidewall capacitance (C_F) due to overlap of the gate with the isolation oxide. This increases the overall gate capacitance [22]. Overall gate capacitance is therefore given by $C_T = C_{\mathrm{ox}} W + 2C_F$, which is greater than C_{ox} given in (10). Hence, the overall V_{th} reduces as shown in Fig. 11 [22]. A much more complex behavior can be observed in the case of trench-isolated buried channel P-MOSFETs, where reduction of the width first decreases the V_{th} until the width is 0.4 μm. The width reduction below 0.4 μm causes a sharp increase in Fig. 12 [27].

4) Effect of Channel Length and V_{th} Rolloff: Threshold voltage of MOSFET decreases as the channel length is reduced. This reduction of threshold voltage with reduction of channel length is known as V_{th} rolloff. Fig. 13 shows the reduction of threshold voltage with reduction in channel length. The principal reason behind this effect is the presence of 2-D field patterns in short-channel devices instead of one-dimensional (1-D) field patterns in long-channel devices. This 2-D field pattern originates from the proximity of source and drain regions [28]. There are depletion regions surrounding the source and drain junctions. In long-channel devices, since the source and drain are far apart, their depletion regions do not have much effect on the potential

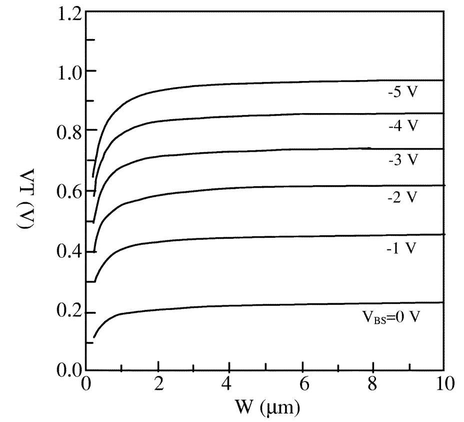


Fig. 11 Variation of threshold voltage with gate width for uniform doping [22].

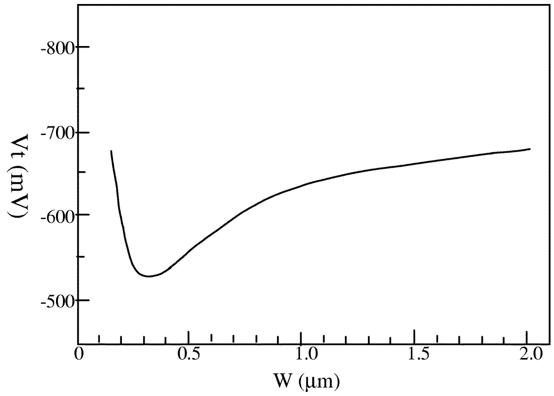


Fig. 12 Variation of threshold voltage with gate width in the case of trench isolated buried channel P-MOSFET showing the anomalous behavior [27].

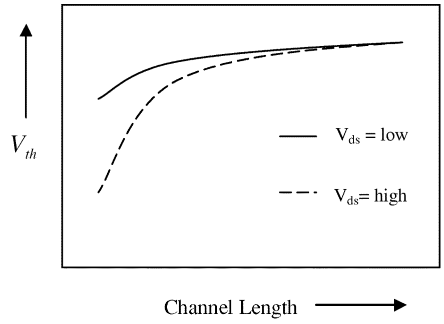


Fig. 13 Threshold voltage rolloff with change in channel length; $V_{\rm th}$ reduction is more severe at higher drain bias.

profile or field pattern in most parts of the channel. However, in the case of short-channel devices, source-to-drain distance is comparable to the depletion width in the vertical direction. As a result, source drain depletion width has a more pronounced effect on potential profiles and field patterns. The source and drain depletion regions now penetrate more into the channel length, resulting in part of the channel being already depleted. Thus, gate voltage has to invert less bulk charge to turn a transistor on (see Fig. 14). In other words, for the same gate voltage, there is more band bending in the Si–SiO$_2$ interface in a short-channel device as compared with a long-channel one. Consequently, the threshold voltage is lower for a short-channel device. The effect of the source drain depletion region is more severe at a high drain bias. High drain bias results in more depletion charge in the channel from the drain and source, resulting in further decrease of the threshold voltage, and hence, larger subthreshold current.

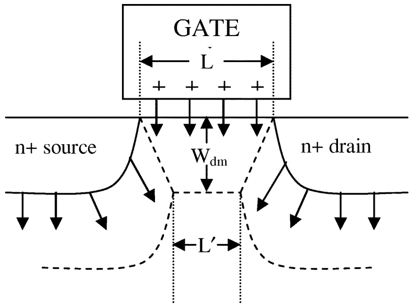
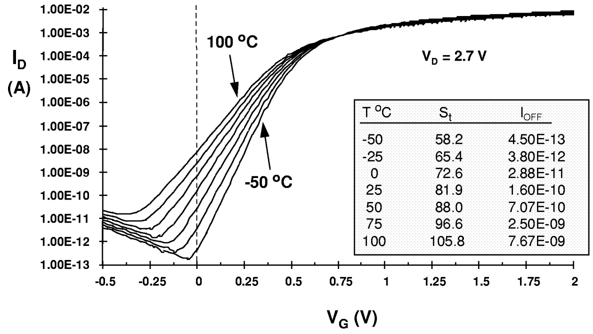


Fig. 14 Schematic diagram for charge-sharing model explaining the reduction of $V_{\rm th}$ due to the source/drain depletion regions. The bulk charge that needs to be inverted is proportional to the area under the trapezoidal region given by $Q'_B \propto W_{\rm dm}(L + L')/2$, which is less than the total depletion charge in the long-channel case, which is $Q_B \propto W_{\rm dm}(L)$ [28].



T °C	S_t	I_{OFF}
-50	58.2	4.50E-13
-25	65.4	3.80E-12
0	72.6	2.88E-11
25	81.9	1.60E-10
50	88.0	7.07E-10
75	96.6	2.50E-09
100	105.8	7.67E-09

Fig. 15 I_D versus V_G showing temperature sensitivity of $I_{\rm OFF}$ [18].

5) Effect of Temperature: Temperature dependence of the subthreshold leakage current is important, since digital very large scale integration (VLSI) circuits usually operate at elevated temperatures due to the power dissipation (heat generation) of the circuit. $\log(I_D)$ versus V_G shows a linear change in subthreshold slope (S_t) with temperature (see Fig. 15) as predicted by the subthreshold current model [15]. In Fig. 15, 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. In this technology, the major component of $I_{\rm OFF}$ is the subthershold leakage; therefore, the temperature dependence of $I_{\rm OFF}$ represents the temperature dependence of the subthreshold leakage. The increase in the $I_{\rm OFF}$ is 0.45–160 pA for the 20-μm-wide device (23 fA/μm to 8 pA/μm). $I_{\rm OFF}$ increases by a factor of 356 for this technology. Two parameters increase the subthreshold leakage as temperature is raised: 1) S_t linearly increases with temperature; and 2) the threshold voltage decreases. The temperature sensitivity of $V_{\rm th}$ was measured to be about 0.8 mV/°C. The temperature sensitivity of $V_{\rm th}$ can be used to estimate $I_{\rm OFF}$ at other temperatures.

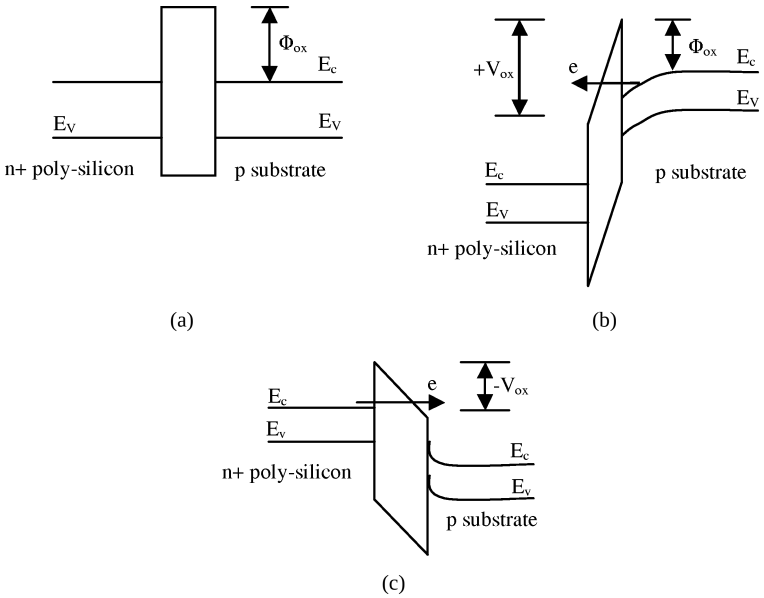


Fig. 16 Tunneling of electrons through an MOS capacitor. (a) Energy-band diagram at flat-band condition. (b) Energy-band diagram with positive gate bias showing tunneling of electron from substrate to gate. (c) Energy-band diagram at negative gate bias showing tunneling of electron from gate to substrate [29].

C. Tunneling into and Through Gate Oxide (I_3)

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 tunneling of electrons from substrate to gate and also from gate to substrate through the gate oxide, resulting in the gate oxide tunneling current.

To understand the phenomenon of tunneling, let us consider an MOS capacitor with a heavily doped n+-type polysilicon gate and a p-type substrate. Also, for simplicity, let us now focus only on the electron tunneling. An energy-band diagram in flat-band condition is shown in Fig. 16(a), where $\Phi_{\rm ox}$ is the Si-SiO$_2$ interface barrier height for electrons. When a positive bias is applied to the gate, the energy-band diagram changes as shown in Fig. 16(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, electrons from the n+ polysilicon can tunnel into or through the oxide layer and give rise to the gate current [see Fig. 16(c)] [29].

The mechanism of tunneling between substrate and gate polysilicon can be primarily divided into two parts, namely: (1) Fowler–Nordheim (FN) tunneling; and (2) direct tunneling. In the case of FN 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 the thickness of the barrier, the barrier height, and the structure of the barrier. Therefore, the tunneling probabilities of a single electron in FN tunneling and direct tunneling are different, resulting in different tunneling currents.

1) Fowler–Nordheim Tunneling: In FN tunneling, electrons tunnel into the conduction band of the oxide layer. Fig. 17 shows the FN tunneling of electrons from the inverted surface to the gate. Ignoring the effect of finite temperature and image-force-induced barrier lowering, the current density in the FN tunneling is given by [29]

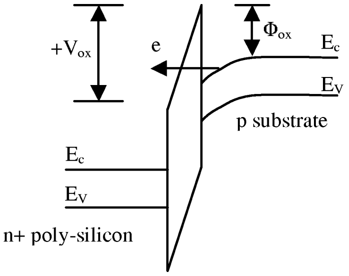


Fig. 17 FN tunneling of electrons.

$$J_{FN} = \frac{q^3 E_{\rm ox}^2}{16\pi^2 \hbar \phi_{\rm ox}} \exp\left(-\frac{4\sqrt{2m^*}\phi_{\rm ox}^{3/2}}{3\hbar q E_{\rm ox}}\right) \qquad (12)$$

where $E_{\rm ox}$ is the field across the oxide; $\phi_{\rm ox}$ is the barrier height for electrons in the conduction band; and m^* is the effective mass of an electron in the conduction band of silicon. The FN current equation represents the tunneling through the triangular potential barrier and is valid for $V_{\rm ox} > \phi_{\rm ox}$, where $V_{\rm ox}$ is the voltage drop across the oxide [30]. The measured value of FN tunneling current is very small; for example, at an oxide field of 8 MV/cm, the FN tunneling current density is about 5×10^{-7} A/cm^2 [29]. Since $\phi_{\rm ox} = 3.1$ eV, short-channel devices mostly operate at $V_{\rm ox} < \phi_{\rm ox}$. Thus, for normal device operation, the FN tunneling current is negligible.

2) Direct Tunneling: In very thin oxide layers (less than 3–4 nm), electrons from the inverted silicon surface, instead of tunneling into the conduction band of SiO$_2$, directly tunnel

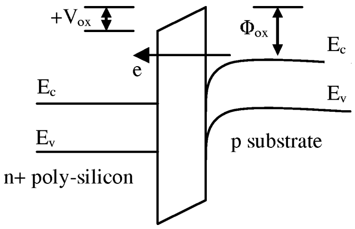


Fig. 18 Direct tunneling of electrons.

to the gate through the forbidden energy gap of the $\mathrm{SiO_2}$ layer [29]. The direct tunneling phenomenon is explained in Fig. 18. In the case of direct tunneling, electrons tunnel through a trapezoidal potential barrier instead of a triangular potential barrier. Hence, the direct tunneling occurs at $V_{\mathrm{ox}} < \phi_{\mathrm{ox}}$ [30]. The equation governing the current density of the direct tunneling is given by [30]

$$J_{\mathrm{DT}} = AE_{\mathrm{ox}}^2 \exp\left\{-\frac{B\left[1-\left(1-\frac{V_{\mathrm{ox}}}{\phi_{\mathrm{ox}}}\right)^{3/2}\right]}{E_{\mathrm{ox}}}\right\} \tag{13}$$

where $A = q^3/16\pi^2\hbar\phi_{\mathrm{ox}}$ and $B = 4\sqrt{2m^*}\phi_{\mathrm{ox}}^{3/2}/3\hbar q$. Direct tunneling current is significant for low oxide thicknesses. Fig. 19 shows the variation of the direct tunneling current density with V_{ox} based on (13).

Potential drop across the oxide is obtained from the fact that applied gate voltage over the flat-band voltage drops across the polysilicon depletion layer, gate oxide, and the rest appear as surface potential.

$$V_{\mathrm{gs}} = V_{\mathrm{fb}} + V_{\mathrm{ox}} + \phi_s + V_{\mathrm{poly}} \tag{14}$$

where V_{gs} is the applied gate bias; ϕ_s is the surface potential; and V_{poly} is the potential drop across the polysilicon depletion region given by $\varepsilon_{\mathrm{ox}}^2 E_{\mathrm{ox}}^2/2q\varepsilon_{\mathrm{si}}N_{\mathrm{poly}}$, where N_{poly} is the doping concentration of polysilicon, $\varepsilon_{\mathrm{si}}$ is the permittivity of silicon, and $\varepsilon_{\mathrm{ox}}$ is the permittivity of $\mathrm{SiO_2}$.

a) 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) [31], [32] (see Fig. 20). 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 positive-channel MOSs (PMOSs), HVB controls the gate to channel leakage in inversion, whereas gate-to-body leakage is controlled by EVB in depletion-inversion and ECB in accumulation [31], [32]. Since the barrier height for HVB (4.5 eV) is considerably higher than barrier height for ECB (3.1 eV), the tunneling current associated with HVB is much less than the current associated with ECB. This results in lower gate leakage current in PMOS than in NMOS [33].

b) 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 the source (I_{gcs}) and the rest goes to the drain (I_{gcd}); and the gate to the substrate leakage current (I_{gb}) (see Fig. 21) [31], [32]. The modeling of each of the components can be found in [31], [32].

c) Effect of quantization of substrate electron energy: Due to high substrate doping level and large electric field at the Si–$\mathrm{SiO_2}$ surface, the quantization of carrier energy occurs within the Si substrate (see Fig. 22). This results in less occupied energy states from which 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 little distance away from the surface and not at the surface as predicted by classical physics. This can be considered as an effective increase in the oxide thickness. Thus, quantization effect modulates the gate direct tunneling current [34].

d) Effect of image-force-induced barrier lowering: The emission of electron from Si to $\mathrm{SiO_2}$ causes build up of image charge at the oxide side of the Si–$\mathrm{SiO_2}$ interface, which results in a reduction in the barrier height at the Si–$\mathrm{SiO_2}$ interface from $\phi_{\mathrm{ox}} = 3.1$ eV by an amount $\Delta\phi$ given by

$$\Delta\phi = \sqrt{\frac{q^3 E_{\mathrm{ox}}}{4\pi\varepsilon_{\mathrm{ox}}}} \tag{15}$$

where $\varepsilon_{\mathrm{ox}}$ is the permittivity of $\mathrm{SiO_2}$. This is called the image-force-induced barrier-lowering effect [29]. Since it modulates the barrier height, it also modulates the gate tunneling current, as the tunneling exponentially depends on ϕ_{ox}.

D. Injection of Hot Carriers from Substrate to Gate Oxide (I_4)

In a short-channel transistor, due to high electric field near the Si–$\mathrm{SiO_2}$ interface, electrons or holes can gain sufficient energy from the electric field to cross the interface potential barrier and enter into the oxide layer (see Fig. 23). This effect is known as hot-carrier injection. The injection from Si to $\mathrm{SiO_2}$ is more likely for electrons than holes, as electrons have a lower effective mass than that of holes, and the barrier height for holes (4.5 eV) is more than that for electrons (3.1 eV) [35].

E. Gate-Induced Drain Leakage (I_5)

GIDL is due to high field effect in the drain junction of an MOS transistor. When the gate is biased to form an accumulation layer at the silicon surface, the silicon surface under the gate has almost same potential as the p-type substrate. Due to presence of accumulated holes at the surface, the surface behaves like a p region more heavily doped than the substrate. This causes the depletion layer at the surface to be much narrower than elsewhere [see Fig. 24(a)]. The narrowing of the depletion layer at or near the surface causes field crowding or an increase in the local electric field, thereby enhancing the high field effects near that region [36]. When the negative gate bias is large (i.e., gate at zero or negative and drain at V_{DD}), the n+ drain region under the gate can be depleted

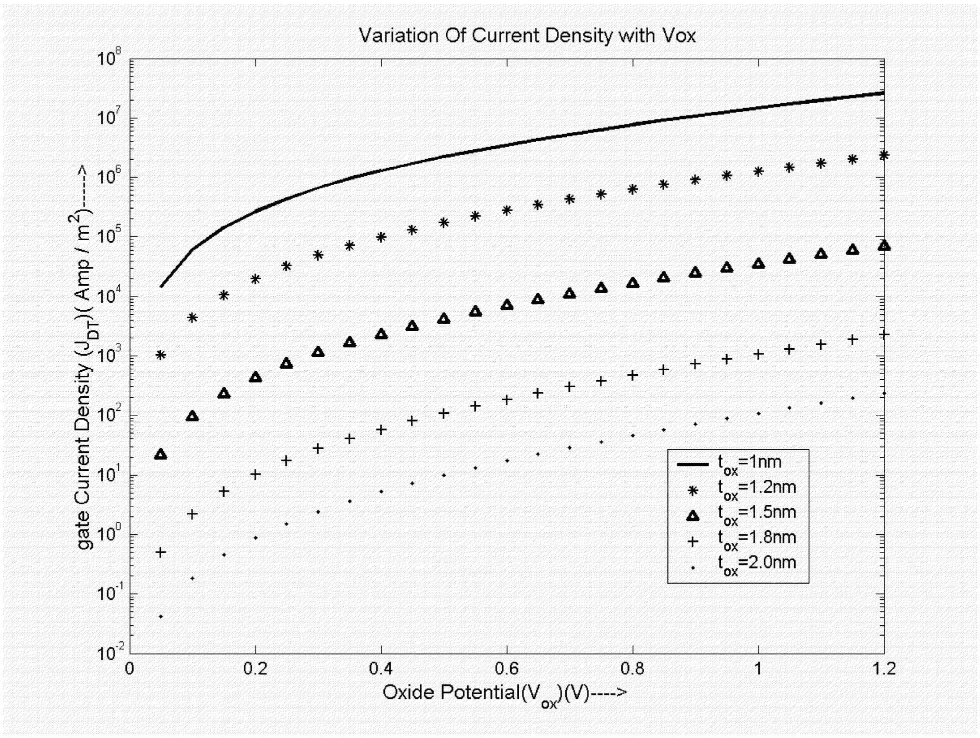


Fig. 19 Simulated direct tunneling current density in thin-oxide polysilicon gate MOS devices.

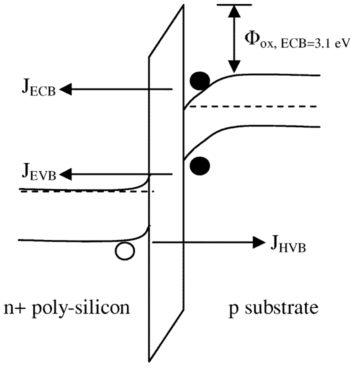


Fig. 20 Three mechanisms for gate leakage [31], [32].

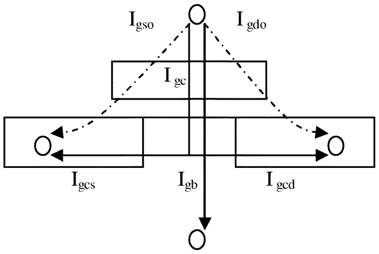


Fig. 21 Components of tunneling current [31], [32].

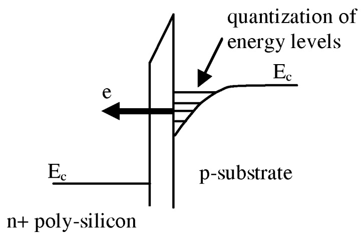


Fig. 22 Quantization of electron energy levels in substrate.

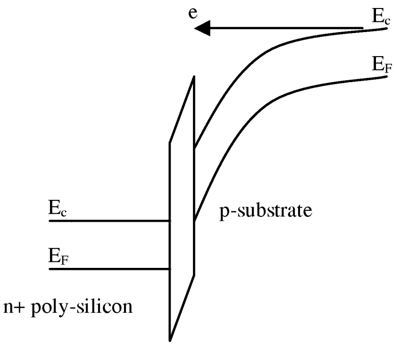


Fig. 23 Injection of hot electrons from substrate to oxide.

and even inverted as shown in Fig. 24(b). This causes more field crowding and peak field increase, resulting in a dramatic increase of high field effects such as avalanche multiplication and BTBT [36]. The possibility of tunneling via near-surface traps also increases. As a result of all 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 GIDL [37]. Thinner oxide thickness and higher $V_{\rm DD}$ (higher potential between gate and drain) enhance the electric field and therefore increase GIDL. The impact of drain and well doping on GIDL is rather complicated. At low drain doping, the electric field is not high enough to cause tunneling. At very high drain doping, the depletion width—and, therefore, the tunneling volume—are limited, causing less

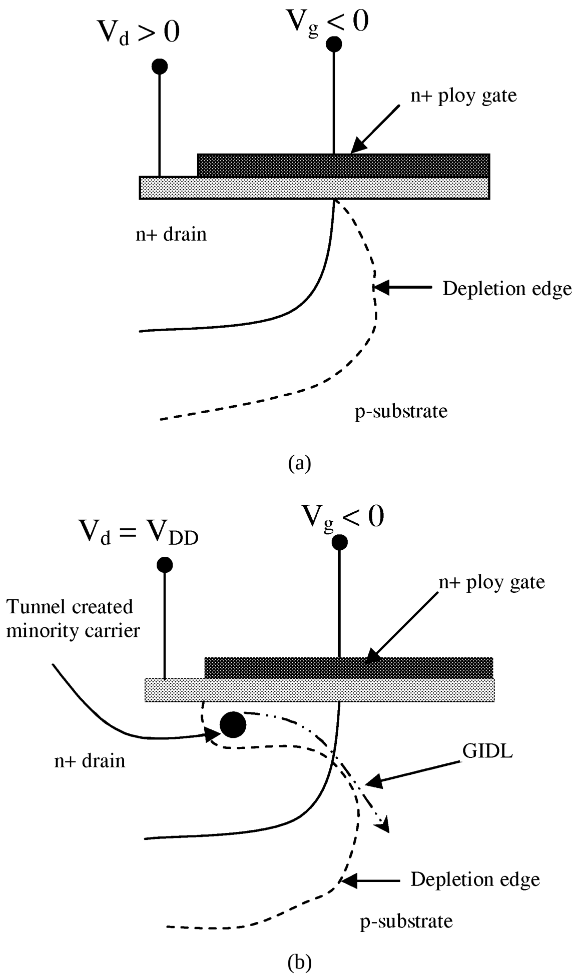


Fig. 24 Condition of the depletion region near the drain-gate overlap region of an MOS transistor when (a) surface is accumulated with low negative gate bias; and (b) n+ region is depleted or inverted with high negative gate bias.

GIDL. Hence, GIDL is worse for moderate drain doping (in between the extremes previously mentioned), where both the electric field and depletion width (tunneling volume) are considerable. Very high and abrupt drain doping is preferred for minimizing GIDL, as it provides lower series resistance required for high transistor drive currents [21].

F. Punchthrough (I_6)

In short-channel devices, due to the proximity of the drain and the source, the depletion regions at the drain-substrate and source-substrate junctions extend into the channel. As the channel length is reduced, if the doping is kept constant, the separation between the depletion region boundaries decreases. An increase in the reverse bias across the junctions (with increase in V_{ds}) also pushes the junctions nearer to each other. When the combination of channel length and reverse bias leads to the merging of the depletion regions, punchthrough is said to have occurred. In submicrometer MOSFETs, a V_{th} adjust implant is used to have a higher doping at the surface than that in the bulk. This causes a greater expansion of the depletion region below the surface (due to smaller doping there) as compared to the surface. Thus, the punchthrough occurs below the surface [38]. An increase in the drain voltage beyond the value required to establish the punchthrough lowers the potential barrier for the majority carriers in the source. Thus, more of these carriers cross the energy barrier and enter into the substrate, and the drain collects some of them. The net effect is an increase in the subthreshold current. Furthermore, punchthrough degrades the subthreshold slope. The device parameter commonly used to characterize the punchthrough is the punchthrough voltage V_{PT}, which estimates the value of V_{ds} for which the punchthrough occurs (i.e., the subthreshold current reaches a particular value) at $V_{\mathrm{gs}} = 0$. It is roughly estimated as the value of the V_{ds} for which the sum of the widths of the drain and source depletion regions is equal to effective channel length [38]

$$V_{\mathrm{PT}} \propto N_B(L - W_j)^3 \qquad (16)$$

where N_B is the doping concentration at the bulk; L is the channel length; and W_j is the junction width.

The most suitable method for controlling the punchthrough 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 implant at the leading edges of the drain and source junctions [38].

III. Leakage Reduction Techniques

For a CMOS circuit, the total power dissipation includes dynamic and static components during the active mode of operation. In the standby mode, the power dissipation is due to the standby leakage current. Dynamic power dissipation consists of two components. One is the switching power due to charging and discharging of load capacitance. The other is short circuit power due to the nonzero rise and fall time of input waveforms. The static power of a CMOS circuit is determined by the leakage current through each transistor. The dynamic (switching) power (P_D) and leakage power (P_{LEAK}) are expressed as

$$P_D = \alpha f C V_{\mathrm{dd}}^2 \qquad (17)$$

$$P_{\mathrm{LEAK}} = I_{\mathrm{LEAK}} \cdot V_{\mathrm{dd}} \qquad (18)$$

where α is the switching activity; f is the operation frequency; C is the load capacitance; V_{dd} is the supply voltage; and I_{LEAK} is the cumulative leakage current due to all the components of the leakage current described in Section II. Due to all the leakage mechanisms described in Section II, leakage current (power) increases dramatically in the scaled devices. Particularly, with reduction of threshold voltage (to achieve high performance), leakage power becomes a significant component of the total power consumption in both active and standby modes of operation (see Fig. 25 [39]). Hence, to suppress the power consumption in low-voltage circuits, it is necessary to reduce the leakage power in both the active and standby modes of operation. The reduction in leakage current has to be achieved using both process- and circuit-level techniques. At the process level, leakage reduction can be achieved by controlling the dimensions (length,

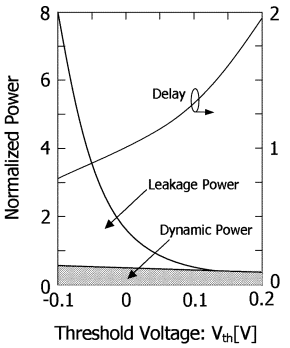


Fig. 25 Power and delay dependence on threshold voltage $(V_{\rm th})$ [39].

oxide thickness, junction depth, etc.) and doping profile in transistors. At the circuit level, threshold voltage and leakage current of transistors can be effectively controlled by controlling the voltages of different device terminals [drain, source, gate, and body (substrate)]. In this section, we first consider major process techniques and then consider several circuit techniques for leakage control and reduction. Though most of the process and circuit techniques described here are used to control the subthreshold leakage, some of them can be used to control other leakage components, too. Reducing all the components of leakage by both process- and circuit-level techniques is of major interest.

A. Channel Engineering for Leakage Reduction

Based on constant field scaling [4], the SCE can be kept under control by scaling down the vertical dimensions, for example, gate insulator thickness, junction depth, along with the horizontal dimensions, while also proportionally decreasing the applied voltages. The substrate doping concentration should increase to decrease the depletion width proportionally. This is shown schematically in Fig. 26 [40]. The principle of constant field scaling lies in scaling the device voltages and the device dimensions (both horizontal and vertical) by the same factor, $K(>1)$, such that the electric field remains unchanged. Constant electric field assures the reliability of the scaled device in terms of hot-carrier injection.

A key parameter is the maximum gate depletion width, $W_{\rm dm}$, within which mobile carriers (holes in the case of nMOSFETs) are swept away by the applied gate field. For uniformly doped case

$$W_{\rm dm} = \sqrt{\frac{4\varepsilon_{\rm si} KT \ln\left(\frac{N_a}{n_i}\right)}{q^2 N_a}} \tag{19}$$

where n_i is the intrinsic carrier concentration. To minimize SCEs, a sufficiently large aspect ratio (AR) of the device is required [41]. AR is defined as

$$\frac{\text{AR} = \text{dimension}_{\text{lateral}}}{\text{dimension}_{\text{vertical}}.} \tag{20}$$

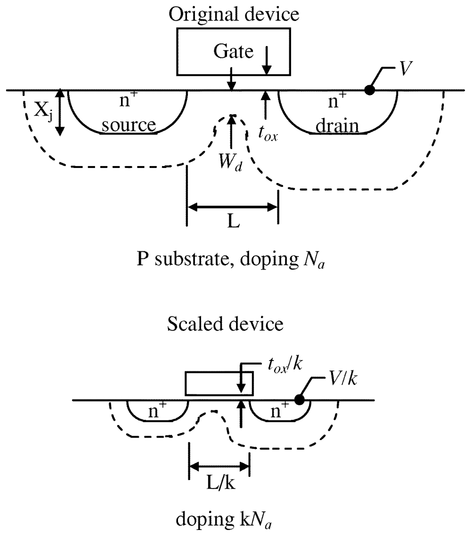


Fig. 26 MOSFET constant-electric-field scaling [40].

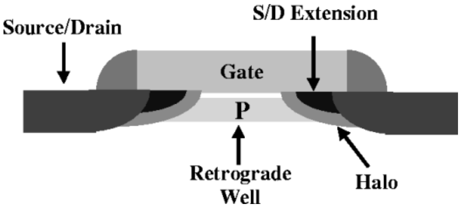


Fig. 27 Graphical representation of different aspects of well engineering [42].

For a MOSFET, AR can be expressed as

$$\text{AR} = \frac{L}{\left[t_{\rm ox}\left(\frac{\varepsilon_{\rm si}}{\varepsilon_{\rm ox}}\right)\right]^{1/3} W_{\rm dm}^{1/3} X_j^{1/3}} \tag{21}$$

where $\epsilon_{\rm si}$ and $\epsilon_{\rm ox}$ are silicon and oxide permittivities; and L, $t_{\rm ox}$, $W_{\rm dm}$, and X_j are channel length, gate oxide thickness, depletion depth, and junction depth, respectively. From (21), we can see that reducing $t_{\rm ox}$, $W_{\rm dm}$, and X_j will reduce the SCE of a MOSFET.

In addition to 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. Supersteep retrograde wells 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 [7]–[10]. Fig. 27 is a schematic representation of the transistor regions that are affected by the different types of well engineering [42]. Retrograde well engineering changes the 1-D characteristics of the well profile by creating a retrograde profile toward the Si–SiO$_2$ surface. The halo profile creates a localized 2-D dopant distribution near the S/D extension regions. The use of these two techniques to increase the device performance, while keeping leakage to a tolerable limit, is discussed in Sections III-A1 and III-A2.

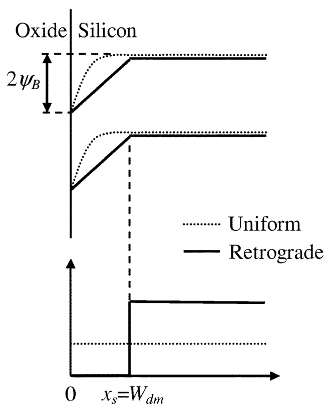


Fig. 28 Band digrams (shown on top) at the threshold condition for a uniformly doped and an extreme retrograde-doped channel (doping profiles shown at bottom) [40].

1) Retrograde Doping: To maintain acceptable OFF-state leakage with continually decreasing channel lengths, both the oxide thickness and the gate-controlled depletion width in silicon

$$W_{\rm dm} = \sqrt{\frac{4\varepsilon_{\rm si}\psi_B}{qN_a}} \tag{22}$$

must be reduced in proportion to the channel length (L) to offset the degradation in SCEs for extremely small devices. This requires an increase in the channel-doping concentration (N_a). This leads to a higher threshold voltage for a uniformly doped channel, according to the following:

$$V_{\rm th} = V_{\rm fb} + 2\psi_B + \frac{\sqrt{4\varepsilon_{\rm si}qN_a\psi_B}}{C_{\rm ox}}. \tag{23}$$

However, if the threshold voltage is not scaled, the device performance for low supply voltages will degrade due to the large reduction in gate drive. To reduce the gate-controlled depletion width while fulfilling the $V_{\rm th}$ reduction trend, retrograde doping can be used.

Retrograde channel doping is a vertically nonuniform, low-high channel doping. It is used to improve the SCEs and to increase surface channel mobility by creating a low surface channel concentration followed by a highly doped subsurface region. The low surface concentration increases surface channel mobility by minimizing channel impurity scattering while the highly doped subsurface region acts as a barrier against punchthrough.

Fig. 28 shows a schematic band-bending diagram at the threshold condition of an extreme retrograde profile with an undoped surface layer of thickness x_s. For the same gate depletion width $(W_{\rm dm})$, the surface electric field and the total depletion charge of an extreme retrograde channel is one-half that of a uniformly doped channel. This reduces the threshold voltage and improves mobility.

Retrograde channel doping allows the threshold voltage $(V_{\rm th})$ to be decoupled from the gate-controlled depletion width $(W_{\rm dm})$. However, the body effect coefficient $m = 1 + (\epsilon_{\rm si}/W_{\rm dm})/(\epsilon_{\rm ox}/t_{\rm ox})$ and the subthresheold slope, $(\ln 10)(mKT/q)$, are still coupled to the gate depletion width $W_{\rm dm}$. For a given $t_{\rm ox}$, reduction in $W_{\rm dm}$ improves SCE, but increases substrate sensitivity and subthreshold slope. Since both subthreshold slope, 2.3 mKT/q, and the substrate sensitivity, $dV_{\rm th}/dV_{\rm bs} = m - 1$, degrade with higher m, m should be kept close to 1. A larger m also results in a lower saturation current in the long-channel limit. Typically, it is required to have $m < 1.5$, or $3t_{\rm ox}/W_{\rm dm} < 1/2$. These design considerations are illustrated in Fig. 29 [43]. The lower limit of $t_{\rm ox}$ is imposed by technology constraints to $V_{\rm dd}/E_{\rm ox}^{\rm max}$, where $E_{\rm ox}^{\rm max}$ is the maximum oxide field. For a given L and $V_{\rm dd}$, the allowable parameter space in a $t_{\rm ox} - W_{\rm dm}$ design plane is a triangular area bounded by SCE, oxide field, and subthreshold slope (also substrate sensitivity) requirements.

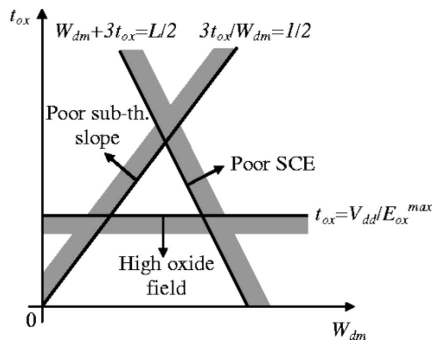


Fig. 29 $t_{\rm ox}$–$W_{\rm dm}$ design space. Some tradeoff among the various factors can be made within the parameter space bounded by SCE, body-effect, and oxide-field considerations [43].

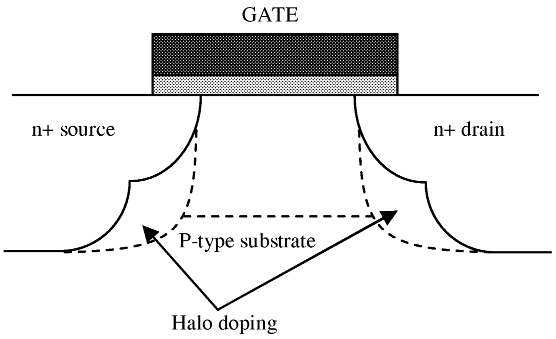


Fig. 30 Halo or nonuniform channel doping.

2) Halo Doping: Halo doping or nonuniform channel profile in a lateral direction was introduced below 0.25-μm technology node to provide another way to control the dependence of threshold voltage on channel length. For n-channel MOSFETs, more highly p-type doped regions are introduced near the two ends of the channel as shown in Fig. 30. Under the edges of the gate, in the vicinity of what will eventually become the end of the channel, point defects are injected during sidewall oxidation. These point defects gather doping impurities from the substrate, thereby increasing the doping concentration near the source and drain end of the channel [44]. More highly doped p-type substrate near the edges of the channel reduces the charge-sharing effects from the source and drain fields, thus reducing the width of the depletion region in the drain-substrate and source-substrate regions. As the channel length is reduced, these highly doped regions consume a larger fraction of

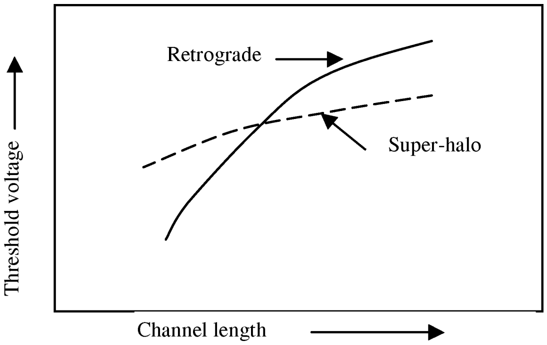


Fig. 31 Short-channel threshold-voltage rolloff for retrograde and superhalo (vertical and lateral nonuniform dopings).

the total channel. Reduction of charge-sharing effects reduces the threshold voltage degradation due to channel length reduction. Thus, threshold voltage dependence on channel length becomes more flat as shown in Fig. 31. Hence, the off-current becomes less sensitive to channel length variation. The reduction in drain and source junction depletion region width also reduces the barrier lowering in the channel, thus reducing DIBL. Since the channel edges are more heavily doped and junction depletion widths are smaller, the distance between source and drain depletion regions is larger. This reduces the punchthrough possibility. The higher doping near the channel edges causes larger BTBT and higher GIDL. The BTBT currents in the high-field region near the drain ultimately limit the halo doping level [40].

B. Circuit Techniques for Leakage Reduction

In this section, four major circuit design techniques—namely, transistor stacking, multiple $V_{\rm th}$, dynamic $V_{\rm th}$, and supply voltage scaling (multiple and dynamic $V_{\rm DD}$) for leakage reduction in digital circuits (logic and memory)—are described.

1) Standby Leakage Control Using Transistor Stacks (Self-Reverse Bias): Subthershold leakage current flowing through a stack of series-connected transistors reduces when more than one transistor in the stack is turned off. This effect is known as the stacking effect. The stacking effect is best understood by considering a two-input NAND gate as shown in Fig. 32. When both M_1 and M_2 are turned off, the voltage at the intermediate node (V_M) is positive due to small drain current [45]. Positive potential at the intermediate node has three effects.

1) Due to the positive source potential V_M, gate-to-source voltage of $M1(V_{\rm gs1})$ becomes negative; hence, the subthreshold current reduces substantially.
2) Due to $V_M > 0$, body-to-source potential $(V_{\rm bs1})$ of M_1 becomes negative, resulting in an increase in the threshold voltage (larger body effect) of M_1, and thus reducing the subthreshold leakage.
3) Due to $V_M > 0$, the drain to source potential $(V_{\rm ds1})$ of M_1 decreases, resulting in an increase in the threshold voltage (less DIBL) of M_1, and thus reducing the subthreshold leakage.

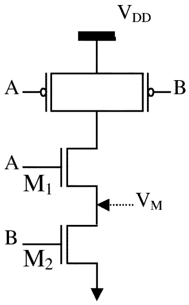


Fig. 32 Stacking effect in two-input NAND gate.

The leakage of a two-transistor stack is an order of magnitude less than the leakage in a single transistor [46]. An analysis of the subthreshold leakage through a stack of n transistor is shown in [47].

Due to the stacking effect, the subthreshold leakage through a logic gate depends on the applied input vector. This makes the total leakage current of a circuit dependent on the states of the primary inputs [48], [49]. The most straightforward way to find a low leakage input vector is to enumerate all combinations of primary inputs. For a circuit with n primary inputs, there are 2^n combinations for input states. Due to the exponential complexity with respect to the number of primary inputs, such an exhaustive method is limited to circuits with a small number of primary inputs. For large circuits, a random search-based technique can be used to find the best input combinations. This method involves generating a large number of primary inputs, evaluating the leakage of each input, and keeping track of the best vector giving the minimal leakage current [48]. A more efficient way is to employ the genetic algorithm to exploit historical information to speculate on new search points with expected improved performance to find a near-optimal solution [47]. The reduction of standby leakage power by application of an input vector is a very effective way of controlling the subthreshold leakage in the standby mode of operation of a circuit. In [50], a stack transistor insertion technique is given. For the gates with high subthreshold leakage in noncritical paths, a leakage control transistor (low $V_{\rm th}$) is inserted in series and is turned off during the standby mode. The technique can effectively reduce the leakage current using single-threshold voltage.

2) Multiple $V_{\rm th}$ Designs: Multiple-threshold CMOS technologies, which provide both high- and low-threshold transistors in a single chip, can be used to deal with the leakage problem. The high-threshold transistors can suppress the subthreshold leakage current, while the low-threshold transistors are used to achieve high performance.

Multiple-threshold voltages can be achieved by the following methods.

1) Multiple channel doping. Multiple-threshold voltages can be achieved by adjusting the channel-doping densities. Fig. 33 shows the threshold voltage at different channel-doping densities [51]. For this approach,

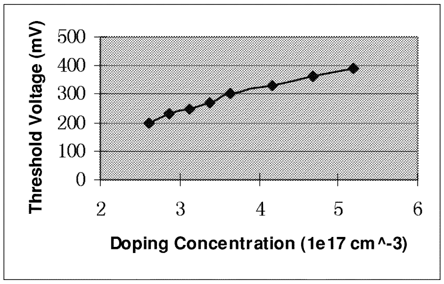


Fig. 33 V_{th} at different channel-doping densities [51].

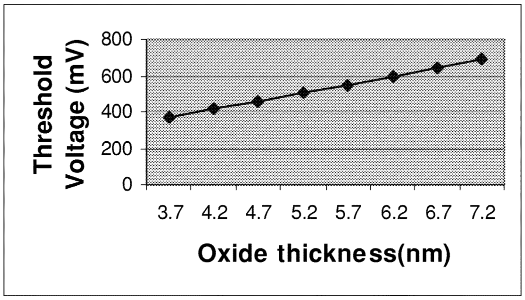


Fig. 34 V_{th} at different oxide thicknesses [51].

two additional masks are required. This technique is commonly used to modify the threshold voltages. However, the threshold voltage can vary due to the nonuniform distribution of the doping density, making it difficult to achieve dual threshold voltages when the threshold voltages are very close to each other.

2) Multiple oxide CMOS ($\mathbf{M_{ox}CMOS}$). Gate oxide thickness can be used to modify the threshold voltage of a transistor. Variation of threshold voltage (V_{th}) with oxide thickness (t_{ox}) for a 0.25-μm device is shown in Fig. 34. Dual V_{th} can be achieved by depositing two different oxide thicknesses. For transistors in noncritical paths, having a higher oxide thickness results in a high threshold voltage, and hence low subthreshold leakage. On the other hand, lower oxide thickness, and hence lower threshold voltage, in critical paths maintains the performance. Higher oxide thickness not only reduces the subthreshold leakage, it also reduces: a) gate oxide tunneling, since the oxide tunneling current exponentially decreases with an increase in the oxide thickness [30]; b) dynamic power consumption, since higher oxide thickness reduces the gate capacitance, which is beneficial for reduction of the dynamic power [51].

For deep-submicrometer devices, increasing the gate oxide thickness has an adverse effect of increasing SCE. To reduce the SCE, the AR of the device must be kept large enough. AR of a device as represented in (21) indicates the short-channel immunity of the transistor—the larger the ratio is, the less the SCEs are [41]. Hence, increased oxide thickness of a transistor should be associated with channel length increase in order to prevent severe SCEs. Fig. 35 shows the relevant channel length for maintaining the AR constant at different gate oxide thicknesses [51]. An advance process technology is required for fabricating M_{OX}CMOS. An algorithm for M_{OX}CMOS design is given in [51].

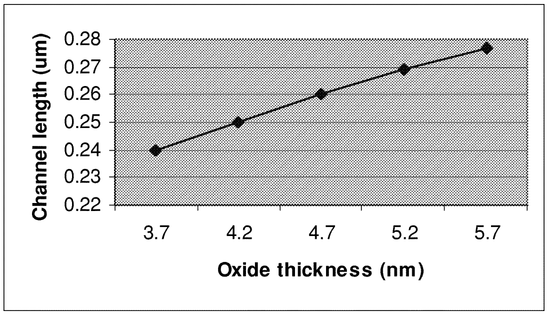


Fig. 35 Channel length at different oxide thicknesses for same AR [51].

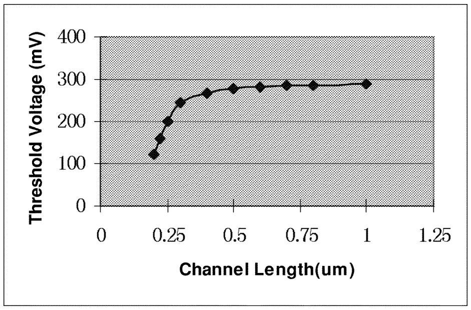
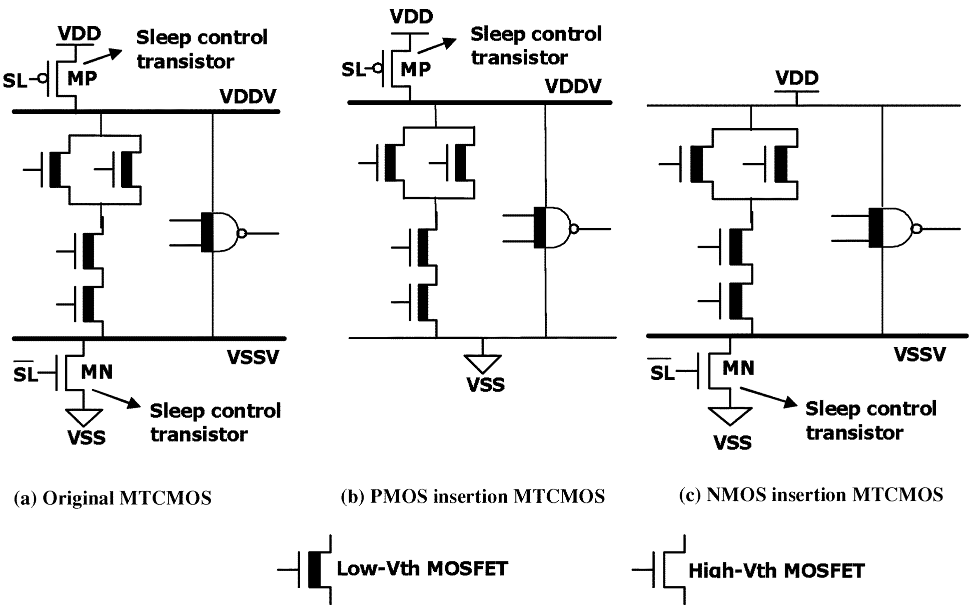


Fig. 36 V_{th} rolloff for NMOS [51].

3) Multiple channel length. For short-channel transistors, the threshold voltage decreases with the decrease in channel length (V_{th} rolloff). Fig. 36 illustrates how scaling of the feature size decreases the threshold voltage based on MINIMOS simulations of 0.25-μm CMOS technology. Hence, different threshold voltages can be achieved by using different channel lengths. Multiple channel length design uses the conventional CMOS technology. However, for the transistors with feature sizes close to 0.1 μm, halo techniques [52] have to be used to suppress the SCE. This causes the V_{th} rolloff to be very sharp; hence, it is nontrivial to control the threshold voltage near the minimum feature size for such technologies. Longer channel lengths for high threshold transistors increase the gate capacitance, which has negative effect on performance and power.

4) Multiple body bias. For bulk silicon devices, the body voltage can be changed to modify the threshold voltage. If separate body biases are applied to different NMOS transistors, the transistors cannot share the same well; therefore, triple well technologies are required. However, it is easier to change the body bias of partially depleted silicon-on-insulator (SOI) devices, since the SOI devices are isolated naturally. For example, consider the double-gate fully depleted (FD) SOI, whose front-gate and back-gate surface potentials are strongly coupled to each other. The threshold voltage can be adjusted by biasing the back-gate voltage.



(a) Original MTCMOS (b) PMOS insertion MTCMOS (c) NMOS insertion MTCMOS

Fig. 37 Schematic of MTCMOS circuit [53].

Based on the multiple-threshold technologies previously described, several multiple-threshold circuit design techniques have been developed recently, as explained in Sections III-B2a–e.

a) Multithreshold-voltage CMOS: Multithreshold-voltage CMOS (MTCMOS) reduces the leakage by inserting high-threshold devices in series to low $V_{\rm th}$ circuitry [53]. Fig. 37(a) shows the schematic of an MTCMOS circuit. A sleep control scheme is introduced for efficient power management. In the active mode, SL is set low and sleep control high $V_{\rm th}$ transistors (MP and MN) are turned on. Since their on-resistances are small, the virtual supply voltages (VDDV and VSSV) almost function as real power lines. In the standby mode, SL is set high, MN and MP are turned off, and the leakage current is low.

In fact, only one type of high $V_{\rm th}$ transistor is enough for leakage control. Fig. 37(b) and (c) shows the PMOS insertion and NMOS insertion schemes, respectively. The NMOS insertion scheme is preferable, since the NMOS on-resistance is smaller at the same width; therefore, it can be sized smaller than corresponding PMOS [54]. MTCMOS can be easily implemented based on existing circuits. A 1-V digital signal processor chip for mobile phone applications has been developed recently [55]. However, MTCMOS can only reduce the standby leakage power, and the large inserted MOSFETs can increase the area and delay. Moreover, if data retention is required in the standby mode, an extra high $V_{\rm th}$ memory circuit is needed to maintain the data [56]. Instead of using high $V_{\rm th}$ sleep control transistors as MTCMOS, super cutoff CMOS (SCCMOS) technique uses low $V_{\rm th}$ transistors with an inserted gate bias generator [57]. For the PMOS (NMOS) insertion, the gate is applied to 0V (VDD) in the active mode, and the virtual VDD (VSS) line is connected to supply VDD (VSS). In the standby mode, the gate is applied to VDD$+\Delta V$ (VSS $-\ \Delta V$) to fully cut off the leakage current. Compared with MTCMOS, SCCMOS circuits can work at lower supply voltages.

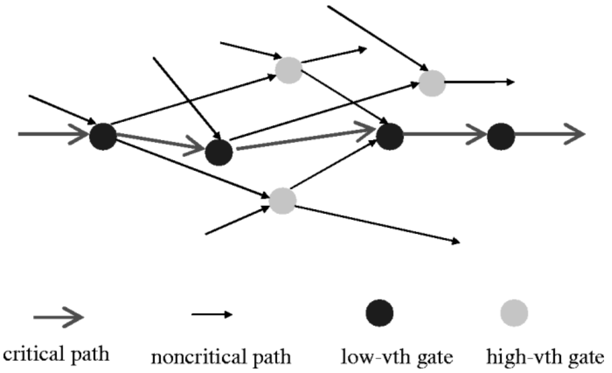


Fig. 38 Dual $V_{\rm th}$ CMOS circuit.

b) Dual threshold CMOS: For a logic circuit, a higher threshold voltage can be assigned to some transistors in noncritical paths so as to reduce the leakage current, while the performance is maintained due to the use of low threshold transistors in the critical path(s) [58], [59]. Therefore, no additional leakage control transistors are required, and both high performance and low power can be achieved simultaneously. Fig. 38 illustrates the basic idea of a dual $V_{\rm th}$ circuit. Fig. 39 shows the path distribution of dual $V_{\rm th}$ and single $V_{\rm th}$ CMOS for a 32-bit adder. Dual $V_{\rm th}$ CMOS has the same critical delay as the single low $V_{\rm th}$ CMOS circuit, but the transistors in noncritical paths can be assigned high $V_{\rm th}$ to reduce leakage power. Dual threshold technique is good for leakage power reduction during both standby and active modes without delay and area overhead.

c) Variable threshold CMOS: Variable threshold CMOS (VTMOS) is a body-biasing design technique [60]. Fig. 40 shows the VTMOS scheme. To achieve different threshold voltages, a self-substrate bias circuit is used to control the body bias. In the active mode, a nearly zero body bias is applied. While in the standby mode, a deeper reverse body bias is applied to increase the threshold voltage and cut off the leakage current. This scheme has been used in a 2-D

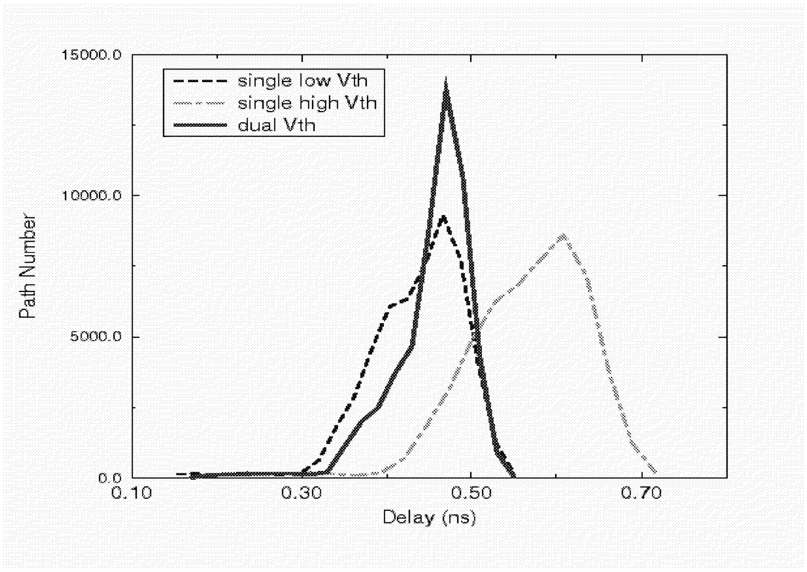


Fig. 39 Path distribution of dual V_{th} and single V_{th} CMOS.

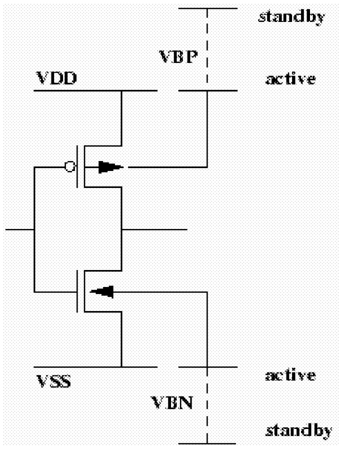


Fig. 40 VTCMOS.

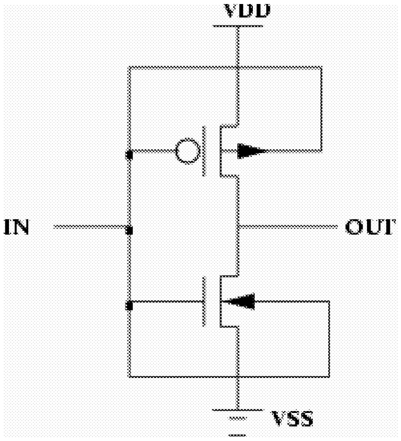


Fig. 41 Schematic of DTMOS inverter.

discrete cosine transform core processor [60]. Furthermore, in the active mode, a slightly forward substrate bias can be used to increase the circuit speed while reducing SCEs [61]. Providing the body potential requires routing the body grid that adds to the overall chip area. Keshavarzi *et al.* reported that reverse body biasing lowers integrated circuit leakage by three orders of magnitude in a 0.35-μm technology [62]. However, more recent data showed that the effectiveness of reverse body bias to lower I_{OFF} decreases as technology scales [62].

d) Dynamic threshold CMOS: For dynamic threshold CMOS (DTMOS), the threshold voltage is altered dynamically to suit the operating state of the circuit. A high threshold voltage in the standby mode gives low leakage current, while a low threshold voltage allows for higher current drives in the active mode of operation. Dynamic threshold CMOS can be achieved by tying the gate and body together [63]. Fig. 41 shows the schematic of a DTMOS inverter. DTMOS can be developed in bulk technologies by using triple wells. "Doping engineering" is needed to reduce the parasitic components [64]. Stronger advantages of DTMOS can be seen in partially depleted SOI devices. Fig. 42 shows the SOI DTMOS structure and layout. In [64], excellent dc inverter characteristics down to 0.2 V and good ring oscillator performance down to 0.3 V are achieved using this method. The supply voltage of DTMOS is limited by the diode built-in potential in bulk silicon technology. The pn diode between source and body should be reverse biased. Hence, this technique is only suitable for ultralow voltage (0.6V and below) circuits in bulk CMOS.

e) Double-gate dynamic threshold SOI CMOS (DGDT-MOS): The double-gate dynamic threshold voltage (DGDT) SOI MOSFET [65] combines the advantages of DTMOS and double-gate FD SOI MOSFETs without any limitation on the supply voltage. Fig. 43 shows the structure of a DGDT SOI MOSFET. A DGDT SOI MOSFET is an asymmetrical double-gate SOI MOSFET. Back-gate oxide is thick enough to make the threshold voltage of the back gate larger than the supply voltage. Since the front-gate and back-gate surface potentials are strongly coupled to each other, the front-gate threshold voltage changes dynamically with the back-gate voltage. Results show that DGDT

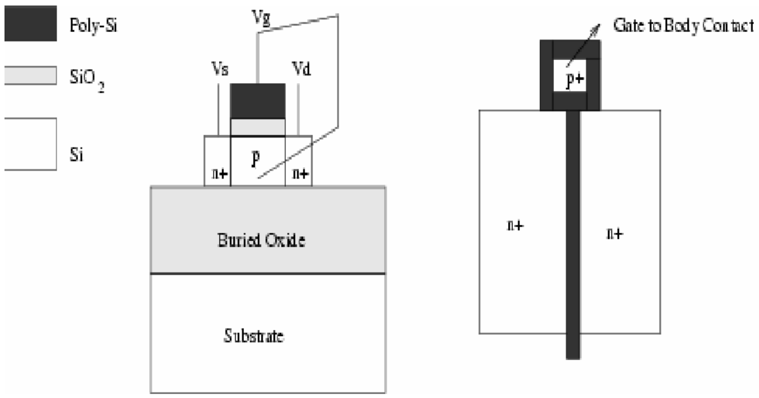


Fig. 42 SOI DTMOS structure and layout.

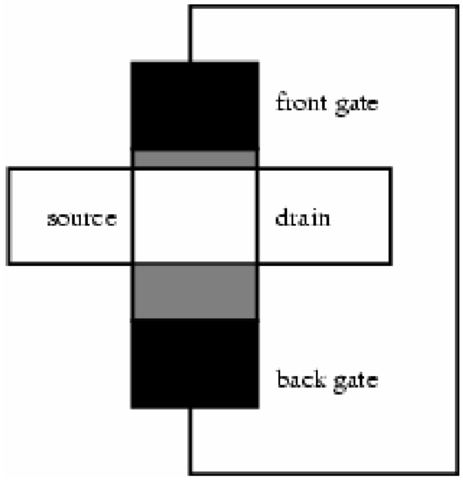


Fig. 43 DGDT SOI MOSFET structure.

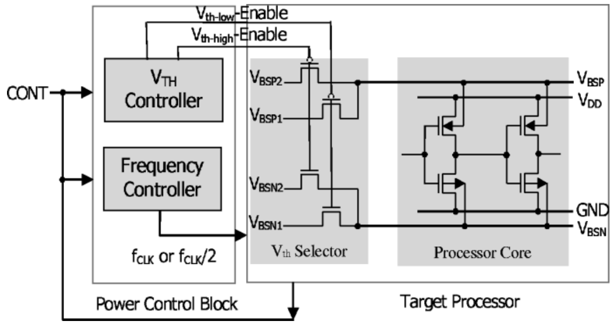


Fig. 44 Schematic diagram of V_{th}-hopping [39].

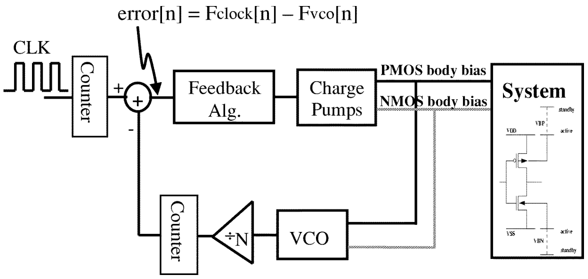


Fig. 45 Schematic of DVTS hardware [68].

SOI MOSFETs have nearly ideal symmetric subthreshold characteristics. Compared with symmetric double-gate SOI CMOS, the power delay product of DGDT SOI CMOS is smaller [66].

3) Dynamic V_{th} Designs: Dynamic threshold voltage scaling is a technique for active leakage power reduction. This scheme utilizes dynamic adjustment of frequency through back-gate bias control depending on the workload of a system. When the workload decreases, less power is consumed by increasing V_{th}. Two varieties of dynamic V_{th} scaling (DVTS) have been proposed, as described later.

a) V_{th}-hopping scheme: Fig. 44 shows the schematic diagram of the V_{th}-hopping scheme [39]. Using the control signal (CONT), which is obtained from software, the power control block generates select signals, $V_{\mathrm{th-low}}$ Enable and $V_{\mathrm{th-high}}$ Enable, which in turn control the substrate bias for the circuit. When the controller asserts $V_{\mathrm{th-low}}$ Enable, V_{th} in the target processor reduces to $V_{\mathrm{th-low}}$. On the other hand, when the controller asserts $V_{\mathrm{th-high}}$ Enable, the target processor V_{th} becomes $V_{\mathrm{th-high}}$. CONT is controlled by software through a software feedback loop scheme [67]. CONT also controls the operation frequency of the target processor. When the controller asserts $V_{\mathrm{th-low}}$ Enable, the frequency controller generates f_{CLK}, and when the controller asserts $V_{\mathrm{th-high}}$ Enable, the frequency controller generates $f_{\mathrm{CLK}}/2$ (say).

b) Dynamic V_{th}-scaling scheme: A block diagram of the DVTS scheme and its feedback loop is presented in Fig. 45 [68]. A clock speed scheduler, which is embedded in the operating system, determines the (reference) clock frequency at run-time. The DVTS controller adjusts the PMOS and NMOS body bias so that the oscillator frequency of the voltage-controlled oscillator tracks the given reference clock frequency. The error signal, which is the difference between the reference clock frequency and the oscillator frequency, is fed into the feedback controller. The continuous feedback loop also compensates for variation in temperature and supply voltage.

4) Supply Voltage Scaling: Supply voltage scaling was originally developed for switching power reduction. It is an effective method for switching power reduction because of the quadratic dependence of the switching power on the

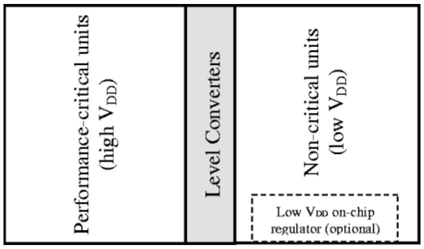

Performance-critical units (high V_{DD})
Level Converters
Non-critical units (low V_{DD})
Low V_{DD} on-chip regulator (optional)

Fig. 46 Two-level multiple supply voltage scheme [75].

supply voltage. Supply voltage scaling also helps reduce leakage power, since the subthreshold leakage due to DIBL decreases as the supply voltage is scaled down [69]. For a 1.2-V 0.13-μm technology, it is shown that the supply voltage scaling has significant impacts on subthreshold leakage and gate leakage (reductions in the orders of V^3 and V^4, respectively) [70].

To achieve low-power benefits without compromising performance, two ways of lowering supply voltage can be employed: static supply scaling and dynamic supply scaling. In static supply scaling, multiple supply voltages are used as shown in Fig. 46. Critical and noncritical paths or units of the design are clustered and powered by higher and lower supply voltages, respectively [71]. Since the speed requirements of the noncritical units are lower than the critical ones, supply voltage of noncritical units can be lowered without degrading system performance. Whenever an output from a low $V_{\rm DD}$ unit has to drive an input of a high $V_{\rm DD}$ unit, a level conversion is needed at the interface [72]. The secondary voltages may be generated off-chip [73] or regulated on-die from the core supply [74]. Dynamic supply scaling overrides the cost of using two supply voltages by adapting the single supply voltage to performance demand. The highest supply voltage delivers the highest performance at the fastest designed frequency of operation. When performance demand is low, supply voltage and clock frequency is lowered, delivering reduced performance but with substantial power reduction [76]. There are three key components for implementing dynamic voltage scaling (DVS) in a general-purpose microprocessor: an operating system that can intelligently determine the processor speed, a regulation loop that can generate the minimum voltage required for the desired speed, and a microprocessor that can operate over a wide voltage range. Fig. 47 shows a DVS system architecture [67]. Control of the processor speed must be under software control, as the hardware alone may not distinguish whether the currently executing instruction is part of a compute-intensive task or a nonspeed-critical task. Supply voltage is controlled by hard-wired frequency-voltage feedback loop, using a ring oscillator as a replica of the critical path. All chips operate at the same clock frequency and same supply voltage, which are generated from the ring oscillator and the regulator.

5) Leakage Reduction Methods for Cache Memory: State-of-the-art microprocessor designs devote a large fraction of the chip area to memory structures, e.g., multiple levels of instruction and data caches, translation look-aside buffers, and prediction tables. For instance, 30% of Alpha 21 264 and 60% of Strong ARM processors are devoted to cache and memory structures [77]. Caches account for a large (if not dominant) component of leakage energy dissipation in recent designs, and will continue to do so in the future. Recent energy estimates for 0.13-μ process technology indicate that leakage energy accounts for 30% of L1 cache energy and as much as 80% of L2 cache energy. To address the problem, several techniques have been proposed in the literature, as explained in Sections III-B5a–c.

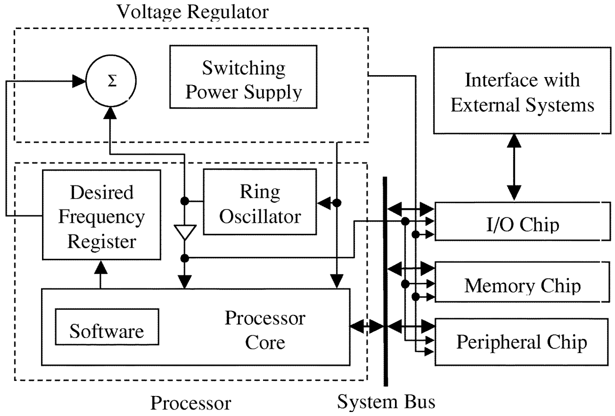


Fig. 47 DVS architecture [67].

a) Data retention gated-ground cache: A data retention gated-ground cache (DRG-Cache) puts the unused portions of the memory core to low leakage mode to reduce power. The key idea is to introduce an extra NMOS transistor (see Fig. 48) in the leakage path from the supply voltage to the ground of the static random access memory (SRAM) cells; the extra transistor is turned on in the used sections and off in the unused sections, essentially "gating" the supply voltage of the cells. Fig. 48 shows the anatomy of the DRG-Cache. Gated ground achieves significantly lower leakage because of the two off transistors connected in series, reducing the leakage current by orders of magnitude; this effect is due to the self-reverse biasing of the stacked transistors, which is called the stacking effect, as described earlier.

Similar to conventional gating techniques, the gated-ground transistor can be shared among multiple SRAM cells from one or more cache blocks. This amortizes the overhead of the extra transistor. Because the size of the gated-ground transistor plays a major role in the data retention capability and stability of the DRG-Cache, and also affects the power and performance savings, the gated-ground transistor must be carefully sized (see Fig. 48) with respect to the SRAM cell transistors. While the gated-ground transistor must be made large enough to sink the current flowing through the SRAM cells during a read/write operation in the active mode and to enhance the data retention capability of the cache in the standby mode, a large gated-ground transistor may reduce the stacking effect, thereby diminishing the energy savings. Moreover, large transistors also increase the area overhead due to gating. In DRG-Cache, the gated-ground transistor is shared by a row of SRAM cells. The gated-ground transistor is controlled by the row decoder logic of the conventional SRAM. The

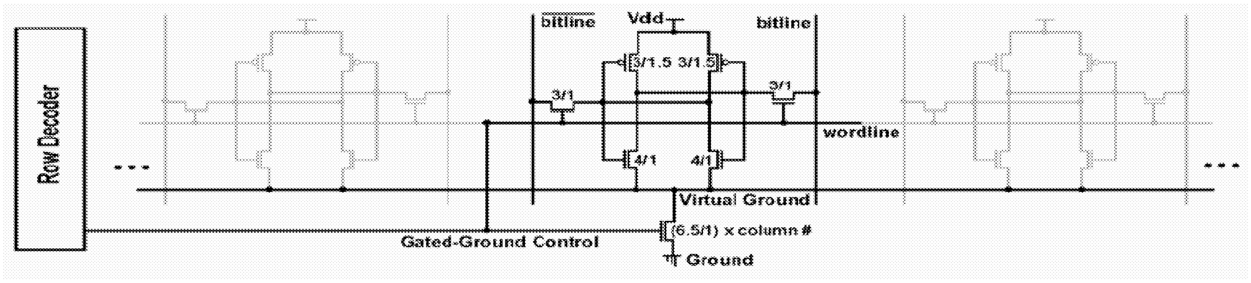


Fig. 48 Anatomy of the DRG-Cache [78].

cells are turned on only when the row is being read from or when data is written into the row. However, this requires the row decoder to drive a larger gate capacitance associated with the gated-ground transistor unlike conventional caches. To maintain performance, proper sizing of the decoder is required.

Conventional SRAM stores the data as long as the power supply is on. This is because the cell storage nodes, which are at zero and one, are firmly strapped to the power rails through conducting devices (by a pulldown NMOS in one inverter and a pull-up PMOS in the other inverter). When the gated-ground transistor is ON, the DRG cache behaves exactly like a conventional SRAM in terms of data storage. Turning off the gated-ground cuts off the leakage path to the ground. However, it also cuts off the opportunity to firmly strap nodes, which are at zero, to the ground. This makes it easier for a noise source to write a one to that node. Turning on the gated-ground transistor restores the zero data. Simulation results show that data is not lost even if the gated-ground transistor is turned off for indefinite time [78].

b) Drowsy cache: Significant leakage reduction can also be achieved by putting the cache into a low-power drowsy mode [79]. In the drowsy mode, the information in the cache line is preserved. However, the line has to be reinstated to a high-power mode before its contents can be accessed. One technique for implementing a drowsy cache is to switch between two different supply voltages in each cache line [79]. Due to SCE in deep-submicrometer devices, subthreshold leakage current reduces significantly with voltage scaling [80]. The combined effect of reduced leakage and supply voltage gives large reduction in the leakage power.

Fig. 49 illustrates the circuit schematic of a SRAM cell connected to the voltage controller. One PMOS pass gate switch supplies the normal supply voltage (VDD) (in the active mode), and the other supplies the low supply voltage (VDDLow) (in the standby mode) for the drowsy cache line. Each pass gate is a high $V_{\rm th}$ device to prevent leakage current from the normal supply to the low supply through the two PMOS pass gate transistors. A separate voltage controller is needed for each cache line. By scaling the voltage of the cells to approximately 1.5 times of $V_{\rm th}$, the state of the memory cell can be maintained. For a typical 70 nm process, the drowsy voltage is about 0.3 V [79]. Since the capacitance of the power rail is very low, the transition time between the high- and low-power state is low. High $V_{\rm th}$ devices are used as the pass transistors that connect the internal inverters of the memory cell to the read/write lines (N1 and N2). This reduces the leakage through the pass transistors, since the read/write lines are maintained in high-power mode.

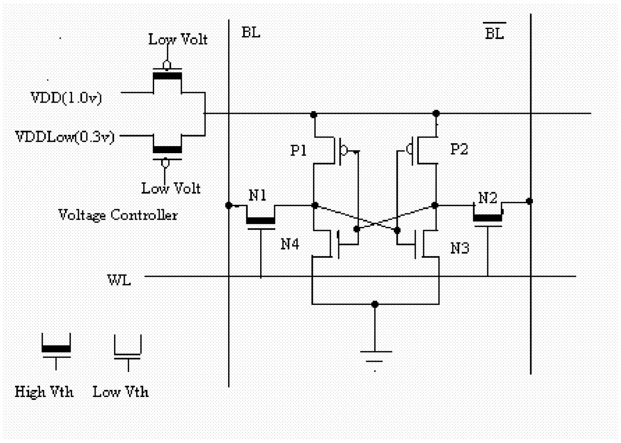


Fig. 49 Schematic of drowsy memory circuit [79].

c) Dynamic threshold voltage $V_{\rm th}$ SRAM: Dynamic $V_{\rm th}$ SRAM (DTSRAM) architecture can be used to reduce leakage energy dissipation in memory structures. Using body biasing, the subthreshold leakage can be reduced without sacrificing data stability [81]. In a time-based dynamic $V_{\rm th}$ scheme, high $V_{\rm th}$ is assigned to the cache lines which are not accessed for a certain period (30 μs–100 μs), and a low $V_{\rm th}$ is assigned to the cache lines which are in frequent use to maintain high performance [82]. Fig. 50 depicts the two dominant leakage paths for a conventional six-transistor SRAM cell, the $V_{\rm dd}$-to-ground and the bit-line-to-ground leakage paths [78]. These two leakage paths make up a high percentage of the total leakage [82]. Fig. 51 shows the schematic of a DTSRAM cache line. The NMOS substrate can be switched to 0 V for high performance. When the cache line is not in use, the substrate can be switched to a negative voltage $(V_{\rm bs})$ to reduce the leakage. Since the transition energy required for a single substrate bias transition is much more than the leakage energy saved during one clock cycle, $V_{\rm th}$ transition cannot be made every clock cycle [82]. Moreover, the performance loss due to negative body bias (i.e., high $V_{\rm th}$) is considerable. To overcome these difficulties, properties of temporal and spatial locality of cache access can be used. In [82], a time-based scheme is described, which instead of turning

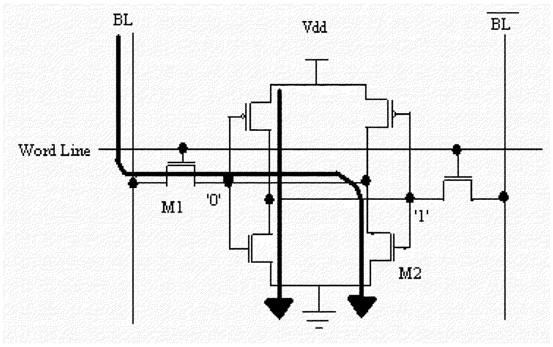


Fig. 50 The two dominant leakage paths ($V_{\rm dd}$ to ground and bitline to ground) for a six-transistor SRAM cell. Leakage through these two paths consist a high percentage of the total leakage [82].

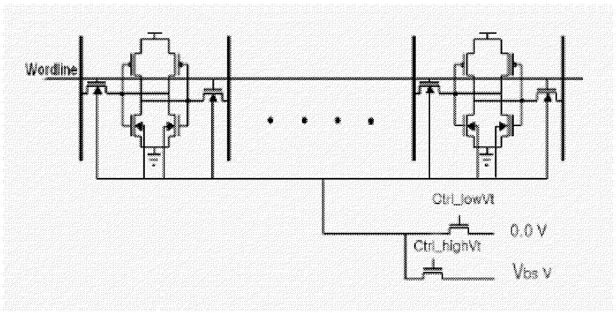


Fig. 51 Schematic of a dynamic $V_{\rm th}$ SRAM set [81].

a cache line to a high $V_{\rm th}$ state right after its access, leaves the cache line in low $V_{\rm th}$ for a certain period (30–100 μs). This ensures that the upcoming accesses within this period will not impose any energy or delay penalties. Moreover, using the spatial locality of program reference, instead of only turning on the accessed cache line, a portion of the cache containing the accessed cache line is turned on. Consequently, subsequent accesses occur in the turned-on portion of the cache. A capacitor-discharging scheme is described in [82] to implement the body-bias control circuit.

IV. Conclusion

With the continuous scaling of CMOS devices, leakage current is becoming a major contributor to the total power consumption. In current deep-submicrometer devices with low threshold voltages, subthreshold and gate leakage have become dominant sources of leakage and are expected to increase with the technology scaling. GIDL and BTBT may also become a concern in advanced CMOS devices. To manage the increasing leakage in deep-submicrometer CMOS circuits, solutions for leakage reduction have to be sought both at the process technology and circuit levels. At the process technology level, well-engineering techniques by retrograde and halo doping are used to reduce leakage and improve short-channel characteristics. At the circuit level, transistor stacking, multiple $V_{\rm th}$, dynamic $V_{\rm th}$, multiple $V_{\rm dd}$, and dynamic $V_{\rm dd}$ techniques can effectively reduce the leakage current in high-performance logic and memory designs.

References

[1] V. De and S. Borkar, "Technology and design challenges for low power and high performance," in *Proc. Int. Symp. Low Power Electronics and Design*, 1999, pp. 163–168.

[2] K. Roy and S. C. Prasad, *Low-Power CMOS VLSI Circuit Design*. New York: Wiley, 2000, ch. 5, pp. 214–219.

[3] C. Mead, "Scaling of MOS technology to submicrometer feature sizes," *Analog Integrated Circuits Signal Process.*, vol. 6, pp. 9–25, 1994.

[4] R. Dennard *et al.*, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. SC-9, p. 256, Oct. 1974.

[5] J. Brews, *High Speed Semiconductor Devices*, S. M. Sze, Ed. New York: Wiley, 1990, ch. 3.

[6] (2001) International Technology Roadmap for Semiconductors. International SEMATECH, Austin, TX. [Online]. Available: http://public.itrs.net/

[7] S. Thompson, P. Packan, and M. Bohr, "Linear versus saturated drive current: Tradeoffs in super steep retrograde well engineering," in *Dig. Tech. Papers Symp. VLSI Technology*, 1996, pp. 154–155.

[8] S. Venkatesan, J. W. Lutze, C. Lage, and W. J. Taylor, "Device drive current degradation observed with retrograde channel profiles," in *Proc. Int. Electron Devices Meeting*, 1995, pp. 419–422.

[9] J. Jacobs and D. Antoniadis, "Channel profile engineering for MOSFET's with 100 nm channel lengths," *IEEE Trans. Electron Devices*, vol. 42, pp. 870–875, May 1995.

[10] W. Yeh and J. Chou, "Optimum halo structure for sub-0.1 μm CMOSFET's," *IEEE Trans. Electron Devices*, vol. 48, pp. 2357–2362, Oct. 2001.

[11] A. Keshavarzi, K. Roy, and C. F. Hawkins, "Intrinsic leakage in low power deep submicron CMOS ics," in *Proc. Int. Test Conf.*, 1997, pp. 146–155.

[12] R. Pierret, *Semiconductor Device Fundamentals*. Reading, MA: Addison-Wesley, 1996, ch. 6, pp. 235–300.

[13] A. S. Grove, *Physics and Technology of Semiconductor Devices*. New York: Wiley, 1967.

[14] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. New York: Cambridge Univ. Press, 1998, ch. 2, pp. 94–95.

[15] ——, *Fundamentals of Modern VLSI Devices*. New York: Cambridge Univ. Press, 1998, ch. 3, pp. 120–128.

[16] J. M. Rabaey, *Digital Integrated Circuits*. Englewood Cliffs, NJ: Prentice-Hall, 1996, ch. 2, pp. 55–56.

[17] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. New York: Cambridge Univ. Press, 1998, ch. 3, pp. 143–144.

[18] 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*, vol. SC-9, p. 256, 1974.

[19] R. Pierret, *Semiconductor Device Fundamentals*. Reading, MA: Addison-Wesley, 1996, ch. 18, pp. 680–681.

[20] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. New York: Cambridge Univ. Press, 1998, ch. 3, p. 130.

[21] V. De, Y. Ye, A. Keshavarzi, S. Narendra, J. Kao, D. Somasekhar, R. Nair, and S. Borkar, "Techniques for leakage power reduction," in *Design of High-Performance Microprocessor Circuits*, A. Chandrakasan, W. Bowhill, and F. Fox, Eds. Piscataway, NJ: IEEE, 2001, ch. 3, pp. 48–52.

[22] S. Chung and C.-T Li, "An analytical threshold-voltage model of trench-isolated MOS devices with nonuniformly doped substrates," *IEEE Trans. Electron Devices*, vol. 39, pp. 614–622, Mar. 1992.

[23] D. Fotty, *MOSFET Modeling with SPICE*. Englewood Cliffs, NJ: Prentice-Hall, 1997, ch. 6, pp. 113–115.

[24] BSIM Group. MOSFET Model. Univ. California, Berkeley. [Online]. Available: http://www-device.eecs.berkeley.edu/~bsim3/

[25] D. Fotty, *MOSFET Modeling with SPICE*. Englewood Cliffs, NJ: Prentice-Hall, 1997, ch. 11, p. 399.

[26] K. Roy and S. C. Prasad, *Low-Power CMOS VLSI Circuit Design*. New York: Wiley, 2000, ch. 2, pp. 26–26.

[27] J. Mandelman and J. Alsmeir, "Anomalous narrow channel effect in trench-isolated burried channel P-Mosfets," *IEEE Electron Device Lett.*, vol. 15, pp. 496–498, Dec. 1994.

[28] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. New York: Cambridge Univ. Press, 1998, ch. 3, pp. 140–143.

[29] ——, *Fundamentals of Modern VLSI Devices*. New York: Cambridge Univ. Press, 1998, ch. 2, pp. 95–97.

[30] K. Schuegraf and C. Hu, "Hole injection Sio2 breakdown model for very low voltage lifetime extrapolation," *IEEE Trans. Electron Devices*, vol. 41, pp. 761–767, May 1994.

[31] BSIM Group. BSIM4.2.1 MOSFET Model. Univ. California, Berkeley. [Online]. Available: http://www-device.eecs.berkeley.edu/~bsim3/

[32] 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," in *Tech. Dig. Int. Electron Devices Meeting*, 2000, pp. 815–818.

[33] F. Hamzaoglu and M. Stan, "Circuit-level techniques to control gate leakage for sub-100 nm CMOS," in *Proc. Int. Symp. Low Power Design*, 2002, pp. 60–63.

[34] 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*, vol. 46, pp. 1464–1471, July 1999.

[35] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. New York: Cambridge Univ. Press, 1998, ch. 2, pp. 97–99.

[36] ——, *Fundamentals of Modern VLSI Devices*. New York: Cambridge Univ. Press, 1998, ch. 2, pp. 99–100.

[37] K. Roy and S. C. Prasad, *Low-Power CMOS VLSI Circuit Design*. New York: Wiley, 2000, ch. 2, pp. 28–29.

[38] ——, *Low-Power CMOS VLSI Circuit Design*. New York: Wiley, 2000, ch. 2, pp. 27–28.

[39] K. Nose, M. Hirabayashi, H. Kawaguchi, S. Lee, and T. Sakurai, "$V_{\rm th}$-Hopping scheme to reduce subthreshold leakage for low-power processors," *IEEE J. Solid-State Circuits*, vol. 37, pp. 413–419, Mar. 2002.

[40] Y. Taur, "CMOS scaling and issues in sub-0.25 μm systems," in *Design of High-Performance Microprocessor Circuits*, A. Chandrakasan, W. J. Bowhill, and F. Fox, Eds. Piscataway, NJ: IEEE, 2001, ch. 2, pp. 27–45.

[41] K. Roy and S. C. Prasad, *Low-Power CMOS VLSI Circuit Design*. New York: Wiley, 2000, ch. 5, pp. 224–226.

[42] S. Thompson, P. Packan, and M. Bohr, "MOS scaling: Transistor challenges for the 21st century," *Intel Technol. J.*, 3rd quarter 1998.

[43] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. New York: Cambridge Univ. Press, 1998, ch. 4, pp. 184–187.

[44] D. Fotty, *MOSFET Modeling with SPICE*. Englewood Cliffs, NJ: Prentice-Hall, 1997, ch. 11, pp. 396–397.

[45] V. De, Y. Ye, A. Keshavarzi, S. Narendra, J. Kao, D. Somasekhar, R. Nair, and S. Borkar, "Techniques for leakage power reduction," in *Design of High-Performance Microprocessor Circuits*, A. Chandrakasan, W. Bowhill, and F. Fox, Eds. Piscataway, NJ: IEEE, 2001, ch. 3, pp. 52–55.

[46] Y. Ye, S. Borkar, and V. De, "New technique for standby leakage reduction in high-performance circuits," in *Dig. Tech. Papers Symp. VLSI Circuits*, 1998, pp. 40–41.

[47] Z. Chen, M. Johnson, L. Wei, and K. Roy, "Estimation of standby leakage power in CMOS circuits considering accurate modeling of transistor stacks," in *Proc. Int. Symp. Low Power Electronics and Design*, 1998, pp. 239–244.

[48] Z. Chen, L. Wei, A. Keshavarzi, and K. Roy, "IDDQ testing for deep submicron ICs: Challenges and Solutions," *IEEE Des. Test Comput.*, pp. 24–33, Mar.–Apr. 2002.

[49] D. Duarte, Y. F. Tsai, N. Vijaykrishnan, and M. J. Irwin, "Evaluating run-time techniques for leakage power reduction," in *Proc. 7th Asia and South Pacific and 15th Int. Conf. VLSI Design*, 2002, pp. 31–38.

[50] M. C. Johnson, D. Somasekhar, and K. Roy, "Leakage control with efficient use of transistor stacks in single threshold CMOS," in *Proc. ACM/IEEE Design Automation Conf.*, 1999, pp. 442–445.

[51] N. Sirisantana, L. Wei, and K. Roy, "High-performance low-power CMOS circuits using multiple channel length and multiple oxide thickness," in *Proc. Int. Conf. Computer Design*, 2000, pp. 227–232.

[52] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. New York: Cambridge Univ. Press, 1998, ch. 4, p. 194.

[53] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, and J. Yamada, "1-V power supply high-speed digital circuit technology with multi-threshold voltage CMOS," *IEEE J. Solid-State Circuits*, vol. 30, pp. 847–854, Aug. 1995.

[54] J. Kao, A. Chandrakasan, and D. Antoniadis, "Transistor sizing issues and tool for multi-threshold CMOS technology," in *Proc. ACM/IEEE Design Automation Conf.*, 1997, pp. 495–500.

[55] S. Mutoh, S. Shigematsu, Y. Matsuya, H. Fukuda, and J. Yamada, "A 1-V multi-threshold voltage CMOS DSP with an efficient power management for mobile phone application," in *Dig. Tech. Papers IEEE Int. Solid-State Circuits Conf.*, 1996, pp. 168–169.

[56] S. Shigematsu, S. Mutoh, Y. Matsuya, Y. Tanabe, and J. Yamada, "A 1-V high speed MTCMOS circuit scheme for power-down applications," *IEEE J. Solid-State Circuits*, vol. 32, pp. 861–869, June 1997.

[57] H. Kawaguchi, K. Nose, and T. Sakurai, "A CMOS scheme for 0.5 V supply voltage with pico-ampere stanby current," in *Dig. Tech. Papers IEEE Int. Solid-State Circuits Conf.*, 1998, pp. 192–193.

[58] L. Wei, Z. Chen, M. Johnson, K. Roy, Y. Ye, and V. De, "Design and optimization of dual threshold circuits for low voltage low power applications," *IEEE Trans. VLSI Systems*, pp. 16–24, Mar. 1999.

[59] P. Pant, V. K. De, and A. Chatterjee, "Simultaneous power supply, threshold voltage, and transistor size optimization for low-power operation of CMOS circuits," *IEEE Trans. VLSI Syst.*, vol. 6, pp. 538–545, Dec. 1998.

[60] T. Kuroda *et al.*, "A 0.9 V 150 MHz 10 mW 4 mm 2-D discrete cosine transform core processor with variable-threshold-voltage scheme," *Dig. Tech. Papers IEEE Int. Solid-State Circuits Conf.*, pp. 166–167, 1996.

[61] Y. Oowaki *et al.*, "A sub-0.1 μm circuit design with substrate-over-biasing," in *Dig. Tech. Papers IEEE Int. Solid-State Circuits Conf.*, 1998, pp. 88–89.

[62] A. Keshavarzi, C. F. Hawkins, K. Roy, and V. De, "Effectiveness of reverse body bias for low power CMOS circuits," in *Proc. 8th NASA Symp. VLSI Design*, 1999, pp. 2.3.1–2.3.9.

[63] F. Assaderaghi, D. Sinitsky, S. Parke, J. Bokor, P. K. Ko, and C. Hu, "A dynamic threshold voltage MOSFET(DTMOS) for ultra-low voltage operation," *Dig. Tech. Papers IEEE Int. Electron Devices Meeting*, pp. 809–812, 1994.

[64] C. Wann, F. Assaderaghi, R. Dennard, C. Hu, G. Shahidi, and Y. Taur, "Channel profile optimization and device design for low-power high-performance dynamic-threshold MOSFET," in *Dig. Tech. Papers IEEE Int. Electron Devices Meeting*, 1996, pp. 113–116.

[65] L. Wei, Z. Chen, and K. Roy, "Double gate dynamic threshold voltage (DGDT) SOI MOSFETS for low power high performance designs," in *Proc. IEEE Int. SOI Conf.*, 1997, pp. 82–83.

[66] ——, "Design and optimization of double-gate fully-depleted SOI MOSFETS for low voltage low power CMOS circuits," *Proc. IEEE Int. SOI Conf.*, pp. 69–70, 1998.

[67] S. Lee and T. Sakurai, "Run-time voltage hopping for low-power real-time systems," *Proc. IEEE/ACM Design Automation Conf.*, pp. 806–809, 2000.

[68] C. H. Kim and K. Roy, "Dynamic $V_{\rm th}$ scaling scheme for active leakage power reduction," in *Proc. Conf. Design, Automation and Test Europe*, 2002, pp. 163–167.

[69] A. J. Bhavnagarwala, B. L. Austin, K. A. Bowman, and J. D. Meindl, "A minimum total power methodology for projecting limits on CMOS GSI," *IEEE Trans. VLSI Syst.*, vol. 8, pp. 235–251, June 2000.

[70] S. Tyagi *et al.*, "A 130 nm generation logic technology featuring 70 nm transistors, dual Vt transistors and 6 layers of Cu interconnects," in *Dig. Tech. Papers Int. Electron Devices Meeting*, 2000, pp. 567–570.

[71] M. Takahashi *et al.*, "A 60-mw MPEG4 video codec using clustered voltage scaling with variable supply-voltage scheme," *IEEE J. Solid-State Circuits*, vol. 33, pp. 1772–1780, Nov. 1998.

[72] Y. Kanno, H. Mizuno, K. Tanaka, and T. Watanabe, "Level converters with high immunity to power-supply bouncing for high-speed Sub-1-V LSI's," in *Dig. Tech. Papers Symp. VLSI Circuits*, 2000, pp. 202–203.

[73] T. Fuse, A. Kameyama, M. Ohta, and K. Ohuchi, "A 0.5 V power-supply scheme for low power LSI's using multi-Vt SOI CMOS technology," in *Dig. Tech. Papers Symp. VLSI Circuits*, 2001, pp. 219–220.

[74] L. R. Carley and A. Aggarwal, "A completely on-chip voltage regulation technique for low power digital circuits," in *Proc. Int. Symp. Low Power Electronics and Design*, 1999, pp. 109–111.

[75] R. K. Krishnamurthy, A. Alvandpour, V. De, and S. Borkar, "High-performance and low-power challenges for sub-70 nm microprocessor circuits," in *Proc. IEEE Custom Integrated Circuits Conf.*, 2002, pp. 125–128.

[76] T. D. Burd, T. A. Pering, A. J. Stratakos, and R. W. Brodersen, "A dynamic voltage scaled microprocessor system," *IEEE J. Solid-State Circuits*, vol. 35, pp. 1571–1580, Nov. 2000.

[77] S. Manne, A. Klauser, and D. Grunwald, “Pipeline gating: Speculation control for energy reduction,” in *Proc. 25th Annu. Int. Symp. Computer Architecture*, 1998, pp. 32–141.

[78] A. Agarwal, H. Li, and K. Roy, “DRG-Cache: A data retention gated-ground cache for low power,” in *Proc. Design Automation Conf.*, 2002, pp. 473–478.

[79] K. Flautner, N. S. Kim, S. Martin, D. Blaauw, and T. Mudge, “Drowsy caches: Simple techniques for reducing leakage power,” in *Proc. 29th Annual Int. Symp. Computer Architecture*, 2002, pp. 148–157.

[80] S. Wolf, *Silicon Processing for the VLSI Era*. Sunset Beach, CA: Lattice Press, 1995, vol. 3, The Submicron MOSFET, pp. 213–222.

[81] H. Mizuno, “An 18 μA (A standby current 1.8-V, 200-MHz microprocessor with self-substrate-biased data-retention mode,” *IEEE J. Solid-State Circuits*, vol. 34, pp. 1492–1500, Nov. 1999.

[82] C. H. Kim and K. Roy, “Dynamic Vt SRAM: A leakage tolerant cache memory for low voltage microprocessors,” presented at the Int. Symp. Low Power Electronics and Design, Monterey, CA, Aug. 2002.

Kaushik Roy (Fellow, IEEE) received the B.Tech. degree in electronics and electrical communications engineering from the Indian Institute of Technology, Kharagpur, India, in 1983 and the Ph.D. degree in electrical and computer engineering department from the University of Illinois, Urbana-Champaign, in 1990.

He was with the Semiconductor Process and Design Center of Texas Instruments, Dallas, TX, where he worked on field programmable gate array architecture development and low-power circuit design. In 1993, he joined the electrical and computer engineering faculty at Purdue University, West Lafayette, IN, where he is currently a Professor. He has published more than 200 papers in refereed journals and conferences, holds five patents, and is the coauthor of *Low-Power CMOS VLSI Design* (New York: Wiley, 1998). His research interests include VLSI design/CAD with particular emphasis in low-power electronics for portable computing and wireless communications, VLSI testing and verification, and reconfigurable computing.

Dr. Roy received the National Science Foundation Career Development Award in 1995, the IBM Faculty Partnership Award, the ATT/Lucent Foundation Award, and Best Paper Awards at the 1997 International Test Conference and the 2000 International Symposium on Quality of IC Design, and is currently a Purdue University Faculty Scholar Professor. He is on the editorial boards of IEEE Design and Test of Computers, IEEE Transactions on Circuits and Systems, and IEEE Transactions on Very Large Scale Integration (VLSI) Systems. He was also Guest Editor for the special issues on low-power VLSI of IEEE Design and Test of Computers (1994) and IEEE Transactions on Very Large Scale Integration (VLSI) Systems (June 2000).

Saibal Mukhopadhyay (Student Member, IEEE) received the B.E. degree in electronics and telecommunication engineering from Jadavpur University, Calcutta, India, in 2000. He is working toward the Ph.D. degree in electrical engineering at Purdue University, West Lafayette, IN.

His research interests include low-power high-performance digital design using deep-submicrometer CMOS, and circuit design with nanodevices.

Hamid Mahmoodi-Meimand (Student Member, IEEE) received the B.S. degree in electrical engineering from Iran University of Science and Technology, Tehran, Iran, in 1998 and the M.S. degree in electrical engineering from the University of Tehran, Tehran, in 2000. He is working toward the Ph.D. degree in electrical engineering at Purdue University, West Lafayette, IN.

His research interests include low-power and high-performance circuit design for deep-submicrometer CMOS technologies.