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].

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

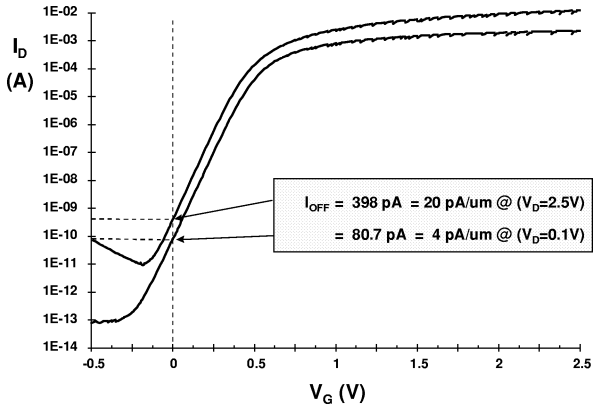
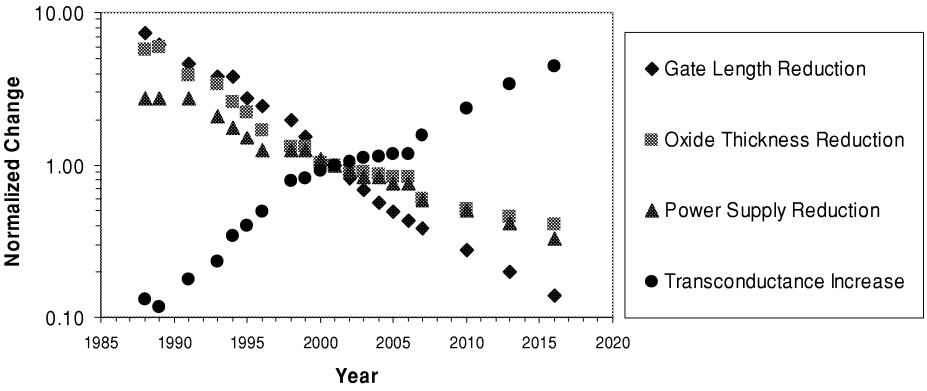
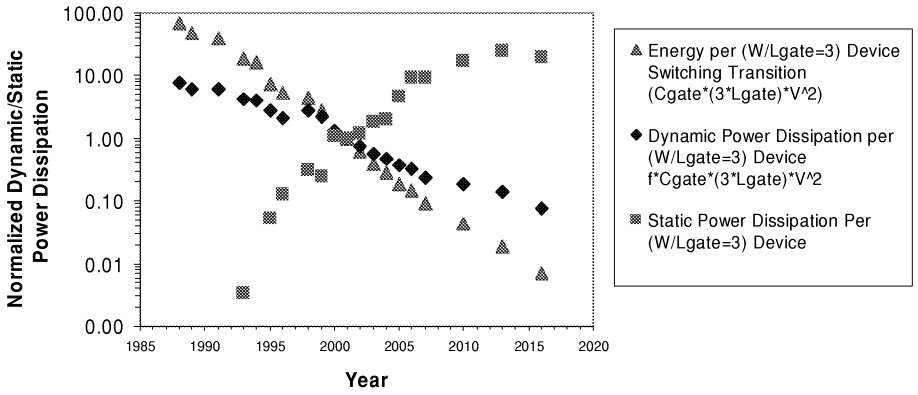


Fig. 1 Log (I_D) versus V_G at two different drain voltages for 20 $\times$ 0.4-μm n-channel transistor in a 0.35-μm CMOS process [2].

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/μm 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



(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 $V_{\rm th}$. 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 $I_{\rm OFF}$ 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

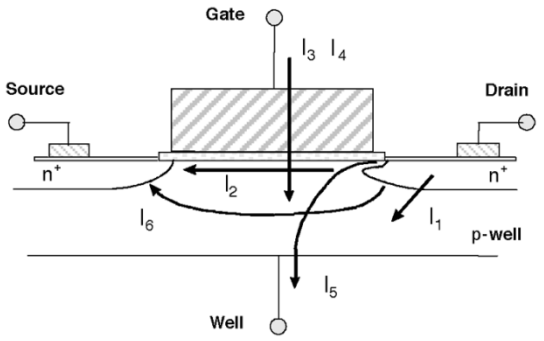


Fig. 3 Summary of leakage current mechanisms of deep-submicrometer transistors.

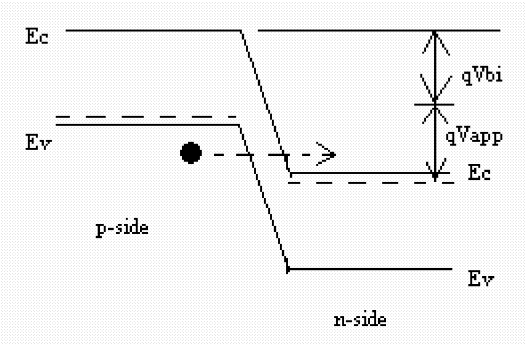


Fig. 4 BTBT in reverse-biased pn junction [14].

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.

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]

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

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]

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

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.

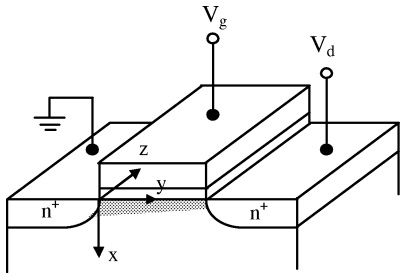


Fig. 5 Variation of minority carrier concentration in the channel of a MOSFET biased in the weak inversion.

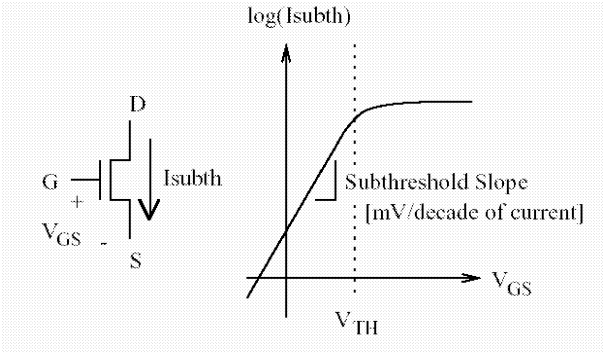


Fig. 6 Subthreshold leakage in a negative-channel metal–oxide–semiconductor (NMOS) transistor.

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.

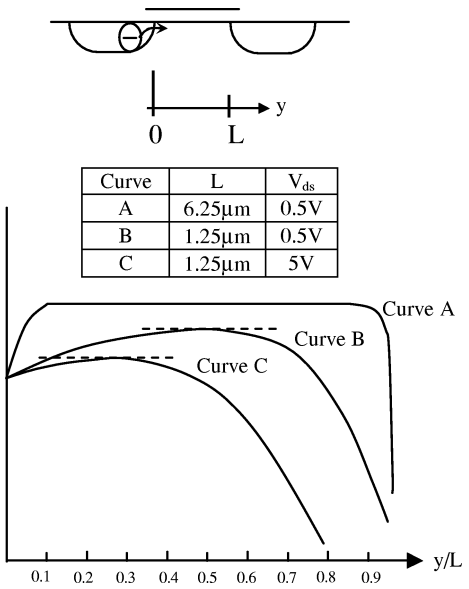
In long-channel devices, the subthreshold current is independent of the drain voltage for $V_{\rm DS}$ larger than a few v_T. On the other hand, the dependence on the gate voltage is exponential, as illustrated in Fig. 6 [16]. The inverse of the slope of the $\log_{10}(I_{\rm ds})$ versus $V_{\rm gs}$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

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

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.

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].

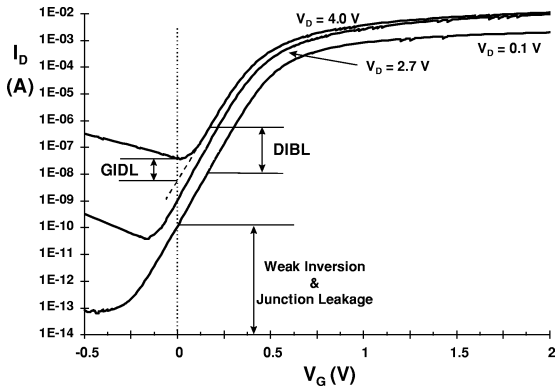


Fig. 8 n channel I_D vs. V_G showing DIBL, GIDL, weak inversion, and pn junction reverse-bias leakage components [11].

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]

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

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

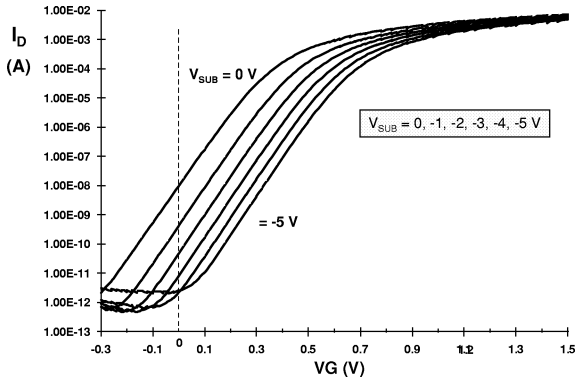


Fig. 9 n channel $\log(I_D)$ versus V_G for six substrate biases on a 0.35-μm logic process technology (V_D = 2.7 V) [11].

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.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.