Physics of Semiconductor Devices

Third Edition

S. M. Sze
National Chiao Tung University
Hsinchu, Taiwan
and
Stanford University
Stanford, California

Kwok K. Ng
Semiconductor Research Corporation
Durham, North Carolina

A JOHN WILEY & SONS, INC., PUBLICATION

TSMC Exhibit 1031

Description of cover photograph.

A scanning electron micrograph of an array of the floating-gate nonvolatile semiconductor memory (NVSM) magnified 100,000 times. NVSM was invented at Bell Telephone Laboratories in 1967. There are more NVSM cells produced annually in the world than any other semiconductor device and, for that matter, any other human-made item. For a discussion of this device, see Chapter 6. Photo courtesy of Macronix International Company, Hsinchu, Taiwan, ROC.

Copyright © 2007 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permission.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic format. For information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data is available.

ISBN-13: 978-0-471-14323-9
ISBN-10: 0-471-14323-5

Printed in the United States of America.

10 9 8 7 6 5 4 3 2

TSMC Exhibit 1031

$[\varepsilon_s/(W_{Dn} + W_{Dp})]$, and C_{ox} should be replaced with C_{ox} in series with the surface depletion capacitance ε_s/W_{Ds}. These substitutions give an expression of[39]

$$S = (\ln 10)\frac{kT}{q}\left[1 + \frac{\varepsilon_{ox}W_{Ds} + \varepsilon_s d}{\varepsilon_{ox}(W_{Dn} + W_{Dp})}\right], \quad (91)$$

where all depletion layers W_{Ds}, W_{Dn}, and W_{Dp} correspond to the condition at threshold ($V_G = V_T$). The subthreshold swing is usually larger than that of conventional surface-channel devices.

The buried-channel device is expected to have higher carrier mobility than surface-channel devices since carriers are free of surface scattering and other surface effects. They are also less affected by the short-channel effects (to be discussed next) such as hot-carrier-induced reliability problems. On the other hand, since the net distance between the gate and the channel is further away and is gate-bias dependent, the transconductance is smaller and variable. Note that if the gate is replaced by a Schottky junction or a *p-n* junction, the device become a MESFET or a JFET correspondingly, both to be discussed in the next chapter.

6.4 DEVICE SCALING AND SHORT-CHANNEL EFFECTS

Since 1959, the beginning of the integrated-circuit era, the minimum feature length has been reduced by more than two orders of magnitude. We expect the minimum dimension will continue to shrink in the foreseeable future, as illustrated in Fig. 1. As the MOSFET dimensions shrink, they need to be designed properly to preserve the long-channel behavior as much as possible. As the channel length decreases, the depletion widths of the source and drain become comparable to the channel length and punch-through between the drain and source will eventually occur. This requires higher channel doping. A higher channel doping will increase the threshold voltage, and in order to control a reasonable threshold voltage, a thinner oxide is necessary. One sees that the device parameters are interrelated, and certain scaling rules are used to optimize the device performance.

Even with the best scaling rules, as the channel length is reduced, departures from long-channel behavior are inevitable. These departures, the short-channel effects, arise as results of a two-dimensional potential distribution and high electric fields in the channel region. The potential distribution in the channel now depends on both the transverse field $\mathscr{E}_x$ (controlled by the gate voltage and the back-substrate bias) and the longitudinal field $\mathscr{E}_y$ (controlled by the drain bias). In other words, the potential distribution becomes two-dimensional, and the gradual-channel approximation (that is, $\mathscr{E}_x \gg \mathscr{E}_y$) is no longer valid. This two-dimensional potential results in many forms of undesirable electrical behavior.

As the electric field is increased, the channel mobility becomes field-dependent, and eventually velocity saturation occurs. (The mobility behavior was discussed in Section 6.2.5.) When the field is increased further, carrier multiplication near the drain occurs, leading to substrate current and parasitic bipolar-transistor action. High

TSMC Exhibit 1031

Table 2 MOSFET Scaling

Parameter	Constant-$\mathscr{E}$	Constant-V	Quasi-constant-V	Actual	Limitation
L	$1/\kappa$	$1/\kappa$	$1/\kappa$	$1/\kappa$	
$\mathscr{E}$	1	> 1	> 1	> 1	
d	$1/\kappa$	$1/\kappa'$	$1/\kappa$	$> 1/\kappa$	Tunneling, defects
r_j	$1/\kappa$	$> 1/\kappa$	$> 1/\kappa$	$> 1/\kappa$	Resistance
V_T	$1/\kappa$	1	$1/\kappa'$	$\gg 1/\kappa$	Off current
V_D	$1/\kappa$	1	$1/\kappa'$	$\gg 1/\kappa$	System and V_T
N_A	κ	κ	κ	$< \kappa$	Junction breakdown

In ideal constant-field scaling parameters are scaled by the same factor. In reality the scaling factors are limited by other reasons and skewed. $1 < \kappa' < \kappa$.

fields also cause hot-carrier injection into the oxide leading to oxide charging and subsequent threshold-voltage shift and transconductance degradation.

These aforementioned phenomena will cause short-channel effects which can be summarized as follows: (1) V_T is not constant with L, (2) I_D does not saturate with V_D bias, both above and below threshold; (3) I_D is not proportional to $1/L$; and (4) device characteristics degrade with operation time. Because short-channel effects complicate device operation and degrade device performance, these effects should be eliminated or minimized so that a physical short-channel device can preserve the *electrical* long-channel behavior. In this section, we discuss MOSFET scaling and the short-channel effects that accompany device miniaturization. [Item-(3) is related to high-field mobility or velocity saturation, and has already been treated in Section 6.2.2.]

6.4.1 Device Scaling

The most-ideal scaling rule to avoid short-channel effects is simply to scale down all dimensions and voltages of a long-channel MOSFET so that the internal electric fields are kept the same.[41] This constant-field scaling is shown in Table 2 and Fig. 25.This approach offers a conceptually simple picture for device miniaturization. All dimensions, including channel length and width, oxide thickness, and junction

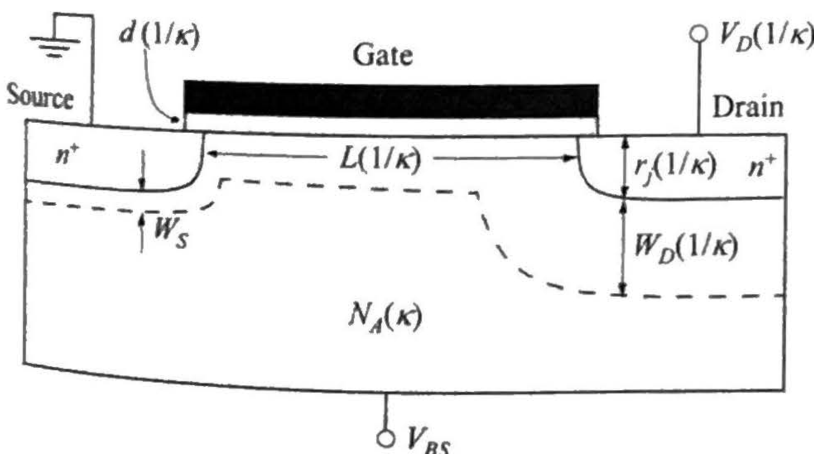

Fig. 25 Physical parameters for MOSFET scaling. Scaling factors for constant-field are indicated.

TSMC Exhibit 1031

depth, are shrunk by the same *scaling factor* κ. The doping level is increased by κ, and all voltages are reduced by κ, leading to a reduction of the junction depletion width by about κ. Note that the subthreshold swing S remains essentially the same since S is proportional to $1 + C_D/C_{ox}$ and both capacitances are scaled up by the same factor κ.

Unfortunately such an ideal scaling rule is hindered by other factors that are fundamentally not scalable. First, the junction built-in voltage and the surface potential for the onset of weak inversion do not scale (only $\approx$ 10% change for 10 times increase in dopings). The range of gate voltage between depletion and strong inversion is approximately 0.5 V. These limitations stem from the fact that both the energy gap and thermal energy kT remain constant. The gate oxide thickness has the technological difficulty of defects as it approaches the low-nm scale. Tunneling through the oxide is another fundamental limitation. The quantum-mechanical effect discussed in Section 4.3.6 degrades the gate capacitance due to the fact that carriers are located at a finite distance ($\approx$1 nm) from the interface. The source and drain series resistance increases when r_j is decreased. This is especially detrimental when the current of the device increases at the same time. The channel doping cannot be increased indefinitely due to *p-n* junction breakdown. The threshold voltage cannot be scaled due to the off-current consideration, even with a fixed subthreshold swing. The supply voltage has been historically slow in scaling because of the system consideration, and also the push for higher speed. These limitations in scaling are summarized in Table 2, resulting in the actual nonideal scaling factors which are shown relative to the constant-field scaling. With these limitations, the field is no longer kept the same, and it increases with smaller gate lengths.

With the above practical limitations, other scaling rules have been proposed. These include constant-voltage scaling,[42] quasi-constant-voltage scaling,[43] and generalized scaling.[44] One other scaling rule with a unique feature of having flexible scaling factors has been proposed.[45] This allows the various device parameters to be adjusted independently as long as the overall behavior is preserved. Therefore, all device parameters do not have to be scaled by the same factor κ. The expression for minimum channel length for which long-channel behavior can be observed is found to follow a simple empirical relation:[45]

$$L \geq C_1[r_j d(W_S + W_D)^2]^{1/3} \tag{92}$$

where C_1 is a constant, and $W_S + W_D$ is the sum of the source and drain depletion widths in a one-dimensional abrupt junction formulation:

$$W_D = \sqrt{\frac{2\varepsilon_s}{qN_A}(V_D + \psi_{bi} - V_{BS})}\,. \tag{93}$$

For $V_D = 0$, W_D equals W_S. A variation of this rule is also presented in Ref. 46.

We have discussed the nonideal factors that hinder constant-field scaling, resulting in some form of a penalty. On the positive notes, there are a couple of disruptive technologies on the horizon that will help scaling. First, MOSFETs built on a three-dimensional structure with an ultra-thin body will effectively eliminate most

TSMC Exhibit 1031

the conduction path for punch-through, and the requirement on channel doping can be relaxed (see Section 6.5.5). Second, research activities looking for gate dielectrics with high dielectric constants have been intense. Such a high-K gate dielectric can relax the physical thickness, improving the defect density and reducing the field for tunnelling. Both of these technologies can help to circumvent or delay the short-channel effects for a particular generation of channel length.

6.4.2 Charge Sharing from Source/Drain

Analysis of the channel charge so far is 1-dimensional, that is, both the inversion charge and depletion charge is completely balanced by the charge on the gate, so they can be treated as charge density. Detailed 2-dimensional examination at the ends of the channel reveals that some of the depletion charge is balanced by the n^+-source and drain, as shown in Fig. 26a. Departure from long-channel behavior can be shown to happen by applying the charge conservation principle to the region bounded by the gate, the channel, and the source/drain,[47]

$$Q'_M + Q'_n + Q'_B = 0\,, \tag{94}$$

where Q'_M is the total charge on the gate, Q'_n is the total inversion-layer charge, and Q'_B is the total ionized impurity in the depletion region. This of course assumes that all oxide and interface charges are zero. The threshold voltage, which can be viewed as voltage required to deplete the total bulk charge Q'_B in the maximum depletion width, is given by

$$V_T = V_{FB} + 2\psi_B + \frac{Q'_B}{C_{ox}A}\,, \tag{95}$$

where A is the gate area $Z{\times}L$. For long-channel devices, $Q'_B = qAN_AW_{Dm}$, where W_{Dm} is the maximum depletion-layer width,

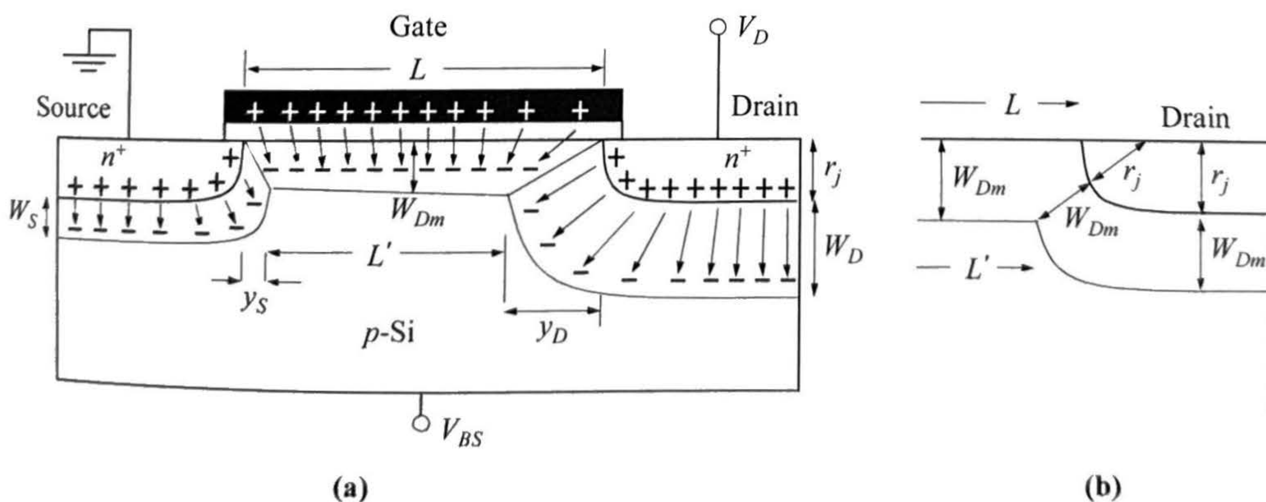

Fig. 26 Charge-conservation model for (a) $V_D > 0$, and (b) $V_D = 0$ where $W_D \approx W_S \approx W_{Dm}$. (After Ref. 47.)

TSMC Exhibit 1031

$$W_{Dm} = \sqrt{\frac{2\varepsilon_s(2\psi_B - V_{BS})}{qN_A}},$$

and 1-D analysis is sufficient.

For short-channel devices, the full effect of Q'_B on the threshold voltage is reduced, because near the source and drain ends of the channel, some field lines originating from the bulk charges under the channel region terminate at the source and drain instead of the gate (Fig. 26a). Note that the horizontal depletion-layer widths y_S and y_D are smaller than the vertical depletion-layer widths W_S and W_D, respectively, because the transverse field strongly influences the potential distribution at the surface.

First-order estimation of the threshold voltage can be made by considering the charge partition. The total bulk depletion charge can be estimated by the trapezoid

$$Q'_B = ZqN_AW_{Dm}\left(\frac{L+L'}{2}\right). \tag{97}$$

For small drain bias, we can assume that $W_D \approx W_S \approx W_{Dm}$, and by straightforward trigonometric analysis (Fig. 26b),

$$L' = L - 2(\sqrt{r_j^2 + 2W_{Dm}r_j} - r_j). \tag{98}$$

The threshold-voltage shift from long-channel behavior is then given by

$$\Delta V_T = \frac{1}{C_{ox}}\left(\frac{Q'_B}{ZL} - qN_AW_{Dm}\right) = -\frac{qN_AW_{Dm}}{C_{ox}}\left(1 - \frac{L+L'}{2L}\right)$$
$$= -\frac{qN_AW_{Dm}r_j}{C_{ox}L}\left(\sqrt{1 + \frac{2W_{Dm}}{r_j}} - 1\right). \tag{99}$$

The negative sign means V_T is lowered and the transistor is easier to turn on. To take into account the effect of the drain voltage and the substrate bias, Eq. 99 can be modified to read [48]

$$\Delta V_T = -\frac{qN_AW_{Dm}r_j}{2C_{ox}L}\left[\left(\sqrt{1 + \frac{2y_S}{r_j}} - 1\right) + \left(\sqrt{1 + \frac{2y_D}{r_j}} - 1\right)\right], \tag{100}$$

where y_S and y_D are given as

$$y_S \approx \sqrt{\frac{2\varepsilon_s}{qN_A}(\psi_{bi} - \psi_s - V_{BS})}, \tag{101a}$$

$$y_D \approx \sqrt{\frac{2\varepsilon_s}{qN_A}(\psi_{bi} + V_D - \psi_s - V_{BS})}. \tag{101b}$$

Note that the threshold voltage becomes a function of both L and V_D. Figure 27 shows this dependence on both channel length and drain bias.

$$W_{Dm} = \sqrt{\frac{2\varepsilon_s(2\psi_B - V_{BS})}{qN_A}}, \tag{96}$$

and 1-D analysis is sufficient.

For short-channel devices, the full effect of Q'_B on the threshold voltage is reduced, because near the source and drain ends of the channel, some field lines originating from the bulk charges under the channel region terminate at the source or drain instead of the gate (Fig. 26a). Note that the horizontal depletion-layer widths y_S and y_D are smaller than the vertical depletion-layer widths W_S and W_D, respectively, because the transverse field strongly influences the potential distribution at the surface.

First-order estimation of the threshold voltage can be made by considering the charge partition. The total bulk depletion charge can be estimated by the trapezoid [47]

$$Q'_B = ZqN_AW_{Dm}\left(\frac{L+L'}{2}\right). \tag{97}$$

For small drain bias, we can assume that $W_D \approx W_S \approx W_{Dm}$, and by straightforward trigonometric analysis (Fig. 26b),

$$L' = L - 2(\sqrt{r_j^2 + 2W_{Dm}r_j} - r_j). \tag{98}$$

The threshold-voltage shift from long-channel behavior is then given by

$$\Delta V_T = \frac{1}{C_{ox}}\left(\frac{Q'_B}{ZL} - qN_AW_{Dm}\right) = -\frac{qN_AW_{Dm}}{C_{ox}}\left(1 - \frac{L+L'}{2L}\right)$$
$$= -\frac{qN_AW_{Dm}r_j}{C_{ox}L}\left(\sqrt{1 + \frac{2W_{Dm}}{r_j}} - 1\right). \tag{99}$$

The negative sign means V_T is lowered and the transistor is easier to turn on. To take into account the effect of the drain voltage and the substrate bias, Eq. 99 can be modified to read [48]

$$\Delta V_T = -\frac{qN_AW_{Dm}r_j}{2C_{ox}L}\left[\left(\sqrt{1 + \frac{2y_S}{r_j}} - 1\right) + \left(\sqrt{1 + \frac{2y_D}{r_j}} - 1\right)\right], \tag{100}$$

where y_S and y_D are given as

$$y_S \approx \sqrt{\frac{2\varepsilon_s}{qN_A}(\psi_{bi} - \psi_s - V_{BS})}, \tag{101a}$$

$$y_D \approx \sqrt{\frac{2\varepsilon_s}{qN_A}(\psi_{bi} + V_D - \psi_s - V_{BS})}. \tag{101b}$$

Note that the threshold voltage becomes a function of both L and V_D. Figure 27 shows this dependence on both channel length and drain bias.

TSMC Exhibit 1031

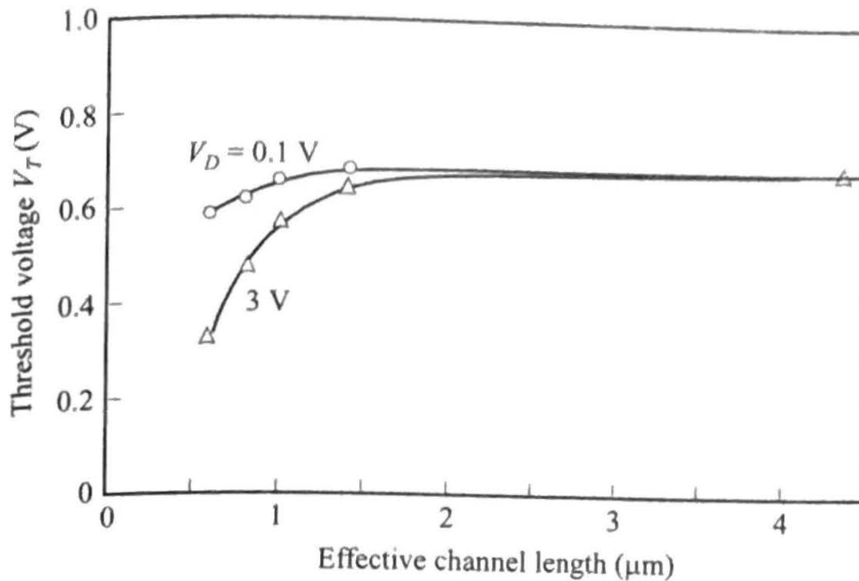

Fig. 27 Dependence of threshold voltage on channel length and drain bias. (After Ref. 49.)

6.4.3 Channel-Length Modulation

Figure 26a also shows that y_D is a high-field region where carriers are swept out efficiently. y_S is a transitional region where the electron concentration is higher than that in the main channel. So for consideration of the channel drift region, the *effective* channel length is more meaningful, given by

$$L_{\text{eff}} = L' = L - y_S - y_D. \tag{102}$$

This factor contributes to a drain-bias dependence of the effective channel length and partially accounts for the nonsaturating current with drain bias. Nevertheless, the change of channel length affects the current only linearly, whereas the barrier lowering caused by the drain bias, considered next, is much more pronounced since the current has an exponential dependence on the barrier.

6.4.4 Drain-Induced Barrier Lowering (DIBL)

We have pointed out that when the source and drain depletion regions are a substantial fraction of the channel length, short-channel effects start to occur. In extreme cases when the sum of these depletion widths approaches the channel length ($y_S + y_D \approx L$), more-serious effects will happen. This condition is commonly called punch-through. The net result is a large leakage current between the source and drain, and that this current is a strong function of the drain bias.

The origin of punch-through is the lowering of the barrier near the source, commonly referred to as DIBL (drain-induced barrier lowering). When the drain is close to the source, the drain bias can influence the barrier at the source end, such that the channel carrier concentration at that location is no longer fixed. This is demonstrated by the energy bands along the semiconductor surface, shown in Fig. 28. For a long-channel device, a drain bias can change the effective channel length, but the barrier at

TSMC Exhibit 1031

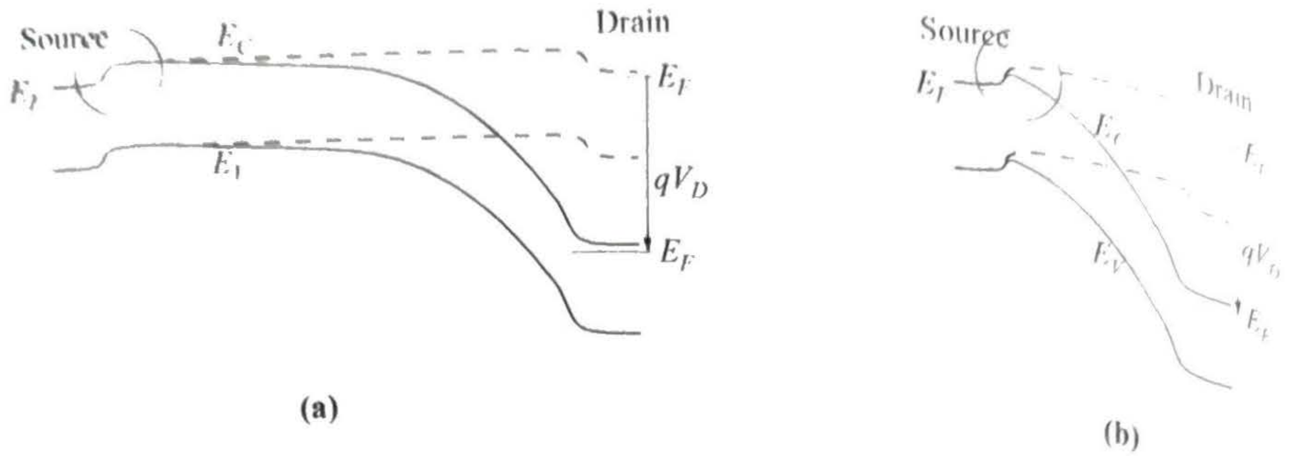

Fig. 28 Energy-band diagram at the semiconductor surface from source to drain, for (a) long-channel and (b) short-channel MOSFETs, showing the DIBL effect in the latter. Dashed lines $V_D = 0$. Solid lines $V_D > 0$.

the source end remains constant. For a short-channel device, this same barrier is no longer fixed. The lowering of the source barrier causes an injection of extra carriers, thereby increasing the current substantially. This increase of current shows up in both above-threshold and subthreshold regimes.

Figure 28 shows that punch-through condition occurs at the semiconductor surface. In practical devices, it is common that the substrate concentration is reduced below the depth of the source/drain junction r_j. A reduced substrate doping widens the depletion widths so punch-through can also happen via a path in the bulk.

An example of severe punch-through characteristics above threshold is shown in Fig. 29a. For this device, at $V_D = 0$ the sum of y_S and y_D is 0.26 μm which is larger than the channel length of 0.23 μm. Therefore, the depletion region of the drain junc-

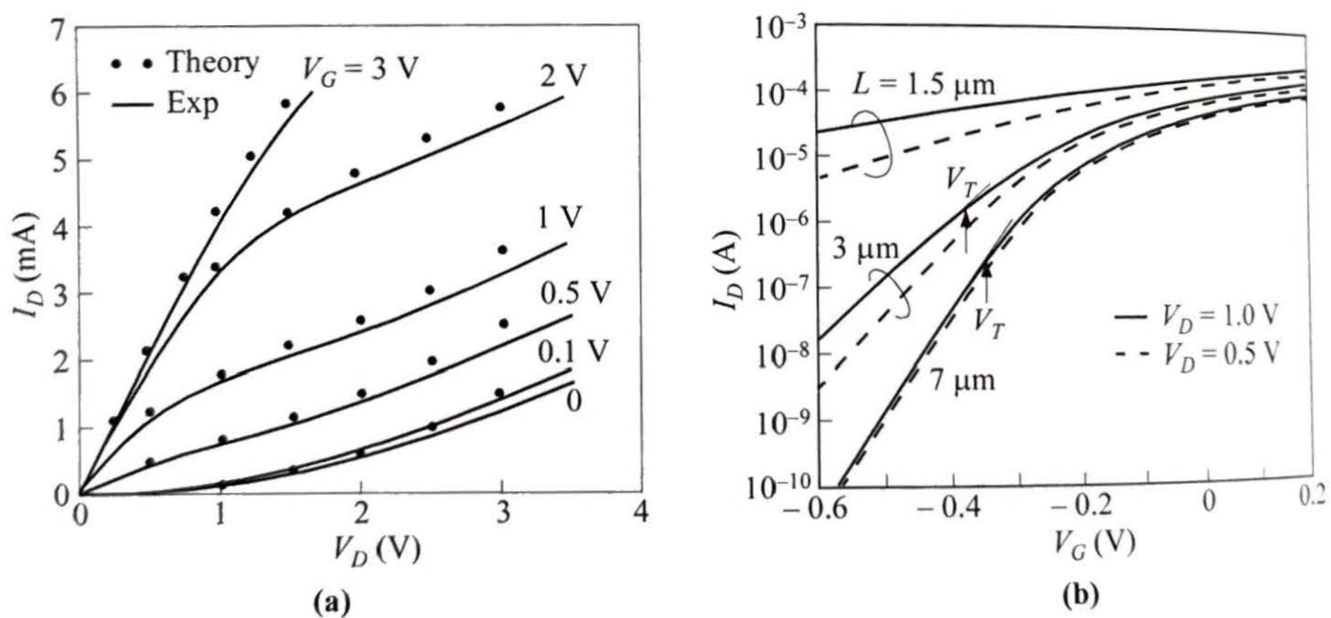

Fig. 29 Drain characteristics of MOSFETs showing DIBL effect. (a) Above threshold. $L =$ 0.23 μm. $d = 25.8$ nm. $N_A = 7\times10^{16}$ cm^{-3}. (b) Below threshold. $d = 13$ nm. $N_A = 10^{14}$ cm^{-3}. (After Ref. 50.)

tion has reached the depletion region of the source junction. Over the drain bias range shown, the device is operated in punch-through condition. Under such a condition, majority carriers in the source region (electrons in this case) can be injected into the depleted channel region, where they will be swept by the field and collected at the drain. The punch-through drain voltage can be estimated by the depletion approximation to be

$$V_{pt} \approx \frac{qN_A(L-y_S)^2}{2\varepsilon_s} - \psi_{bi}. \tag{103}$$

Drain current will be dominated by the space-charge-limited current:

$$I_D \approx \frac{9\varepsilon_s \mu_n A V_D^2}{8L^3} \tag{104}$$

where A is the cross-sectional area of the punch-through path. The space-charge-limited current increases with V_D^2 and is parallel to the inversion-layer current. The calculated points in the figure are from a 2-dimensional computer calculation incorporating the punch-through effect and field-dependent mobility effect.

The DIBL effect on subthreshold current is shown in Fig. 29b, for various channel lengths. The device with a 7-μm channel length shows long-channel behavior, that is, the subthreshold drain current is independent of drain voltage when $V_D > 3kT/q$ (Eq. 57). For $L = 3$ μm, there is a substantial dependence of current on V_D, with a corresponding shift of V_T (which is at the point of current departure of the *I-V* characteristic from the straight line). The subthreshold swing also increases. For an even shorter channel, $L = 1.5$ μm, long-channel behavior is totally lost. The subthreshold swing becomes much worst and the device cannot be turned off any more.

6.4.5 Multiplication and Oxide Reliability

We pointed out earlier that due to nonideal scaling, the internal electric fields in MOSFETs would increase with shorter channel lengths. In this section we discuss the anomalous currents associated with high fields, as well as their impacts. Figure 30 depicts all parasitic currents in addition to the main channel current. Note that the highest field occurs near the drain, and this is the location where most of the anomalous currents originate.

First, as the channel carriers (electrons) go through the high-field region, they acquire extra energy from the field without losing it to the lattice. These energetic carriers are called *hot carriers* whose kinetic energy is measured above the conduction band E_C. This extra energy, if larger than the Si/SiO_2 barrier (3.1 eV), enables them to escape to the oxide layer and to the gate terminal, and gives rise to a gate current.

Another major phenomenon happening in the high-field region is impact ionization which generates extra electron-hole pairs. These extra electrons go directly to the drain and add to the channel current. The path of the generated holes, however, is more diverse. A small fraction of them are driven to the gate, analogous to the hot electrons mentioned before. The majority of the generated holes flow to the substrate. For short-channel devices, some holes will go to the source. The division of these

TSMC Exhibit 1031