



ENCYCLOPEDIA OF

Physical Science AND Technology

THIRD EDITION

Ele - Fi

Volume 5



ENCYCLOPEDIA OF
Physical Science
AND **Technology**

THIRD EDITION

Ele-Fi
Volume 5

Editor-in-Chief

Robert A. Meyers, Ramtech, Inc.



ACADEMIC PRESS

A Harcourt Science and Technology Company

San Diego

San Francisco

Boston

New York

London

Sydney

Tokyo

General Library System
University of Wisconsin - Madison
28 State Street
Madison, WI 53706-1494
U.S.A.

This book is printed on acid-free paper. ∞

Copyright © 2002 by ACADEMIC PRESS

All Rights Reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.

Requests for permission to make copies of any part of the work should be mailed to:
Permissions Department, Harcourt Inc., 6277 Sea Harbor Drive,
Orlando, Florida 32887-6777

Academic Press

A Harcourt Science and Technology Company
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
<http://www.academicpress.com>

Academic Press

Harcourt Place, 32 Jamestown Road, London NW1 7BY, UK
<http://www.academicpress.com>

Library of Congress Catalog Card Number: 2001090661

International Standard Book Number:

0-12-227410-5 (Set)	0-12-227420-2 (Volume 10)
0-12-227411-3 (Volume 1)	0-12-227421-0 (Volume 11)
0-12-227412-1 (Volume 2)	0-12-227422-9 (Volume 12)
0-12-227413-X (Volume 3)	0-12-227423-7 (Volume 13)
0-12-227414-8 (Volume 4)	0-12-227424-5 (Volume 14)
0-12-227415-6 (Volume 5)	0-12-227425-3 (Volume 15)
0-12-227416-4 (Volume 6)	0-12-227426-1 (Volume 16)
0-12-227417-2 (Volume 7)	0-12-227427-X (Volume 17)
0-12-227418-0 (Volume 8)	0-12-227429-6 (Index)
0-12-227419-9 (Volume 9)	

PRINTED IN THE UNITED STATES OF AMERICA

01 02 03 04 05 06 MM 9 8 7 6 5 4 3 2 1



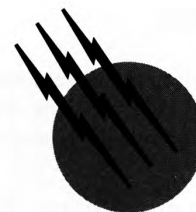
Contents

<i>Foreword</i>	<i>ix</i>
<i>Preface</i>	<i>xiii</i>
<i>Guide to the Encyclopedia</i>	<i>xv</i>

E

Electrical and Electronic Measurements	1
Barrie A. Gregory, Retired	
Electric and Hybrid-Electric Vehicles	39
John S. Reuhl	
Electric Machines, Design	53
Enrico Levi	
Electric Machines, Drives	69
Zivan Zabar	
Electric Machines, Rotating, Construction and Theory	97
Leo Birenbaum and Enrico Levi	
Electric Propulsion	125
Robert G. Jahn and Edgar Y. Choueiri	
Electrochemical Engineering	143
Geoffrey Prentice	
Electrochemistry	161
Donald T. Sawyer	
Electrodynamics, Quantum	199
W. P. Healy	
Electrolyte Solutions, Thermodynamics	219
J. Barthel, W. Kunz, and R. Neueder	
Electrolyte Solutions, Transport Properties	243
J. Barthel, W. Kunz, P. Turq, and O. Bernard	
Electromagnetic Compatibility	261
J. F. Dawson, A. C. Marvin, and C. A. Marshman	
Electromagnetics	277
Sheila Prasad	
Electron Microprobe Analysis of Minerals	293
Peter Schiffman and Sarah Roeske	
Electrons in Solids	307
Richard H. Bube	
Electron Spin Resonance	331
Larry Kevan	
Electron Transfer Reactions	347
Gilbert P. Haight, Jr.	
Electrophoresis	363
S. P. Spragg	

F



Field-Effect Transistors

A. Blicher*

I. H. Kalish

David Sarnoff Research Center

K. F. Brennan

Georgia Institute of Technology

- I. Fabrication Techniques
- II. The Field-Effect
- III. Junction Field-Effect Transistor
- IV. Metal-Insulator-Semiconductor Field-Effect Transistor

GLOSSARY

Abrupt semiconductor junction Junction with a step-like doping density transition from the p - to the n -type semiconductor.

Built-in potential Potential difference existing in a pn junction in equilibrium.

Channel Path for the majority carrier current flow.

Drain Sink for the majority charge carriers.

Flat-band voltage Metal oxide semiconductor transistor gate voltage required to counterbalance the work-function difference between the gate metal and the semiconductor and the effect of all electric charges in the gate dielectric. The higher the flat-band voltage, the higher is the device gate turn-on (threshold) voltage.

Gate Control electrode whose field modulates the magnitude of the current of the field-effect transistor.

Insulated-gate field-effect transistor (IGFET) Field-

effect transistor with the conductive gate insulated from the semiconductor.

Junction field-effect transistor (JFET) Field-effect transistor with a junction-type gate.

Junction depletion region The $p - n$ junction space charge formed at the interface of the p - and n -type semiconductors.

Metal oxide semiconductor field-effect transistor (MOSFET) An IGFET with the metal gate insulated with an oxide from the semiconductor surface; a four-terminal device consisting of the source, drain, gate, and substrate (body).

Pentode Electron tube that consists of five electrodes and whose anode current stops increasing after the anode voltage reaches some critical (saturation) value.

Source Electrode that supplies the majority carriers.

Threshold voltage Turn-on voltage that must be applied to the MOSFET gate in order to induce a strongly conducting channel.

Triode Electron tube consisting of three electrodes. Its

*Formerly with RCA Laboratories

anode current increases with the anode voltage and stops growing only if there is not enough electrons emitted by the cathode.

A FIELD-EFFECT TRANSISTOR (FET) is a semiconductor device in which the magnitude of the majority carrier current depends on the modulation of the conductance by an electric field. A potential applied to a control electrode (gate) produces the field that controls the current flow from the source electrode to the drain electrode. Because of the very high gate input impedance, the input current is very small. Two types of FETs are of significance: JFET with a junction-type gate, and IGFET with an insulated gate. FETs offer several advantages over bipolar devices in many applications due to their high input impedance, their good thermal stability, and the relative ease of integration of many devices on the same semiconductor chip. The last-named feature has been the basis for a rapid expansion in the development of VLSI (very large scale integration).

I. FABRICATION TECHNIQUES

The basic techniques for fabricating FETs include semiconductor crystal oxidation, photolithography, ion implantation, diffusion, epitaxial crystal growth, etching of patterns and surfaces, and metallization.

In many cases *pn* junctions are formed by the diffusion of *n*-type and/or *p*-type impurities. Diffusion involves two steps: gaseous or impurity coating deposition and drive-in diffusion, which moves the impurity atoms below the semiconductor surface. The process takes place in furnaces with a neutral or oxidizing atmosphere and at elevated temperatures. Diffusion usually requires temperatures above 1000°C.

In order to restrict the deposition of impurities to well-defined areas, the semiconductor surface is first oxidized to form a barrier to the impurities. Then, the crystal is covered with a photoresist film, a material sensitive to ultraviolet light. By the use of appropriate masking (photomasks) the photosensitive film is exposed, baked, and chemically etched away in those areas that undergo diffusion. In the case of very small dimensions, ultraviolet light is replaced by X-ray or electron beam exposure.

To fabricate thin, lightly, and uniformly doped layers above a highly doped layer of the same or different polarity, it is advantageous to use epitaxial growth in place of diffusion. The semiconductor substrate (e.g., a silicon wafer) is placed in a chamber on a carbon susceptor, which is raised by induction heating to a very high temperature (> 1000°C). A gas, such as silane (SiH_4), silicon tetrachloride (SiCl_4), or trichlorosilane (SiHCl_3), is passed over the

silicon wafer. In the case of SiH_4 , at temperatures above 1000°C the silane decomposes irreversibly into hydrogen and silicon, the latter of which is deposited as a film on the silicon wafer surface.

Another widely used and more precise technique for the deposition of very thin layers is known as ion implantation. The doping impurity is first ionized in high vacuum and then accelerated by high electric field before reaching the semiconductor substrate. After the removal of the substrate from the vacuum chamber the impurity is diffused and heat-treated (annealed) in order to repair the mechanical damage to the crystal caused by the highly energetic ions. The advantage of ion implantation over thermal deposition and diffusion is due to the introduction of the dopant at low temperature, at specific places, and with a very exact impurity dosage. Masking can be accomplished by using silicon dioxide (SiO_2), silicon nitride (Si_3N_4), metals, or photoresists.

In molecular beam epitaxy both the semiconductor atoms and doping atoms are deposited on heated substrates in an ultrahigh vacuum in order to very slowly grow single crystals. The precision of the deposited film thickness is achieved by either blocking or opening a mechanical shutter placed in the path of the molecular beam. The shutter is controlled by a preprogrammed computer. Layers less than 1 nm thick are achieved by means of this technique. Molecular beam epitaxy is used in the fabrication of modulation-doped FETs.

During device processing, it is necessary to remove the oxide in the areas undergoing diffusion or to open some areas to make electrical contacts. Cleaning of semiconductor surfaces is of great importance, since to a large extent it determines the quality of the device. All these functions can be performed by dry chemical etching, which can be accomplished by gas plasma discharges in a closed chamber. Whereas wet etching tends to undercut photoresist or silicon dioxide films, dry etching, being anisotropic, avoids this undesirable feature.

Ohmic (i.e., very low resistance) electrical contacts to the semiconductor are made by using very carefully selected metals. The most commonly used metals for silicon devices are aluminum-silicon and gold-silicon eutectics that melt at comparatively low temperatures. These metals are first deposited and then alloyed into the semiconductor, forming a good electrical contact with either the highly doped n^+ or the highly doped p^+ silicon.

II. THE FIELD-EFFECT

A. Definition of the Field-Effect

Unlike a metal or an insulator, the electrical conductivity of a semiconductor can be readily engineered. The

conductivity of a semiconductor can be controlled in several different ways through the direct modification of the conductivity by the introduction of impurities called dopants into the host semiconductor lattice; through the generation of carriers by optical excitation; through electrical injection; and by the field-effect. In all of these cases, the conductivity can be altered by many orders of magnitude and many useful electrical devices can be made. The field-effect is arguably the most ubiquitous of these techniques since it underlies the workings of the transistor devices used in virtually all integrated circuits.

The electrical conductivity of a semiconductor can be expressed as the product of the charge density, n , charge, q , and carrier mobility, μ , as

$$\sigma = qn\mu \quad (1)$$

Altering the charge density, n , can control the conductivity of the semiconductor. The field-effect provides a very useful and reversible means of engineering the conductivity of a semiconductor layer. The field-effect can be defined as the modulation of the conductivity of an underlying semiconductor layer by the application of a voltage to a surface electrode, typically called the gate. The application of the gate bias alters the charge concentration within the underlying semiconductor. Depending upon the magnitude and sign of the applied gate voltage, charge can be attracted or repulsed from the gate. As a result, the conductivity of the underlying semiconductor layer can be increased, decreased, or undergo a change in carrier type, i.e., change from n -type (majority carrier electrons) to p -type (majority carrier holes) or from p - to n -type. The semiconductor layer can become highly conductive or resistive thereby altering its current-carrying properties. When the gate is biased such that the semiconductor layer is conducting, a current can flow through the layer producing a conducting channel. Likewise, when the gate is biased such that the semiconductor layer is insulating, the current can be constricted. By controlling the voltage state of the gate, the current flow in the semiconductor layer can be altered.

B. Types of Gating Action

Different types of structures exist that exploit the field-effect. There are three main types. These are the junction field-effect, metal-semiconductor field-effect, and the metal-insulator-semiconductor field-effect structures. In each structure the basic operating principles are the same, i.e., the conductivity of the underlying semiconductor is modulated by the application of a gate bias. The only significant difference is the technique by which the field-effect is produced.

The junction field-effect for an n -channel device utilizes two $p^+ - n$ junctions in opposition. In an

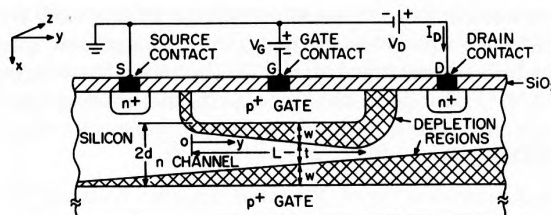


FIGURE 1 Schematic cross section of a JFET.

asymmetrically doped $p^+ - n$ junction the depletion region is mainly formed within the lighter doped n -side of the junction. Application of a reverse bias increases the depletion region width. If the two junctions are placed spatially in opposition, as shown in Fig 1, the two depletion regions extend toward one another constricting the conducting channel formed between them. At sufficiently high reverse bias, the two depletion regions touch and the channel is closed. Hence, the conducting properties of the semiconductor are modulated by the reverse biased diodes that form the gates of the device. Devices that utilize $p - n$ junctions to induce the field effect are called junction field-effect transistors, JFETs.

The field-effect can be induced using a Schottky barrier diode in place of the $p^+ - n$ diodes in JFETs. In this case, a Schottky barrier is formed on top of a thin semiconductor layer. The thin semiconductor layer is in turn grown on top of a semi-insulating material. A conducting channel is formed between the depletion region formed by the Schottky barrier and the semi-insulating material. Reverse biasing of the Schottky barrier further constricts the channel by increasing the depletion region width. In this way, the reverse biased Schottky barrier controls the conducting properties of the layer and can open and close the conducting channel. Devices that include Schottky barriers to induce the field-effect are called metal-semiconductor field-effect transistors, MESFETs.

The last type of field-effect structure discussed here relies on a metal-insulator-semiconductor, MIS, structure. In this structure, an MIS is formed on top of a semiconductor material. Depending upon the bias condition of the MIS, the underlying semiconductor layer can be accumulated, depleted, or inverted. When a voltage of opposite sign to the majority carriers is applied to the metal gate, majority carriers are attracted to the semiconductor-insulator interface. The structure is said to be in accumulation. If a voltage of the same sign as the majority carriers is applied to the metal gate, then the majority carriers are repelled from the interface forming a depletion region. Under these conditions the structure is said to be in depletion and the semiconductor becomes highly resistive. If the voltage is further increased in magnitude, minority carriers of opposite sign to the voltage are

attracted to the insulator-semiconductor interface and the structure is said to be inverted. Devices that operate with an MIS gate are called MISFETs. The most common type of MISFET is made with silicon as the semiconductor and silicon dioxide as the insulator. These structures are called MOSFETs.

III. JUNCTION FIELD-EFFECT TRANSISTOR

A. Basic Structure

For simplicity, our initial discussion of the JFET is limited to the n -channel device operating by drift of electrons. The results apply as well to the p -channel devices, depending on the drift of holes, after the appropriate change of semiconductor and bias polarities.

An n -type semiconductor layer is deposited by epitaxy on a p^+ -semiconductor substrate. Subsequently another p^+ region is deposited on top of the n layer (Fig. 1). Two p^+n gate junctions are thus formed. They are electrically connected (not shown in the figure) and serve as control electrodes (gates). A reverse (negative) potential is applied between the gate and the grounded source electrode S . A positive potential V_D is applied between the drain D and the source. The reverse gate potential creates a depletion region W , thick in the lightly doped n region and very thin (not shown in the figure) in the p^+ region. The variable gate voltage, by changing the depletion region thickness, modulates the magnitude of the current in the channel. The channel thickness $t(y)$ is the difference between the undepleted channel thickness $2d$ and the total depletion region thickness $2W(y)$,

$$t(y) = 2[d - W(y)] \quad (2)$$

1. Modes of Operation and Symbols

Most JFETs operate in the depletion mode; that is, they conduct current (are on) with zero voltage applied to the gate. The normally off (enhancement-type) JFETs do not conduct current for zero gate bias and are suitable for very low power applications. The enhancement mode of operation is made possible by using a very lightly doped channel region so that the built-in junction potential is sufficient to deplete the channel region entirely. A forward voltage (positive for the p^+n junction) greater than the threshold voltage V_T reduces the depletion region thickness and initiates the current flow.

The symbols representing p - and n -channel JFETs in the depletion and enhancement modes of operation are shown in Fig. 2.

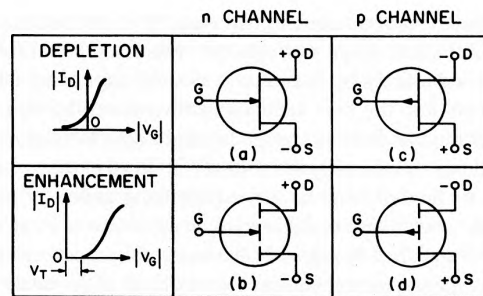


FIGURE 2 Symbols for depletion (normally on) and enhancement (normally off) n - and p -channel JFETs.

B. JFET Static Characteristics

The static (output) characteristics represent the relationship between the drain current I_D and drain-to-source voltage V_D at various values of the gate-to-source voltage V_G (Fig. 3). The study of these characteristics is simplified by considering the case of a uniformly doped channel region and abrupt (steplike) gate junctions.

If the channel length L (Fig. 1) is greater than the maximum channel thickness $2d$, it is reasonable to assume that the channel thickness $2(d - W)$ will vary slowly with distance y and that the field in the channel along y is very small compared with the transverse field in the depletion region. This model is known as the gradual channel approximation (Shockley). For nonuniform doping the result is appreciably different only for very short channels ($L < 2 \mu\text{m}$). The first-order theory also assumes that carrier velocity is field independent.

1. Linear and Saturation Regions

For a uniformly doped channel region the magnitude of the resistive incremental voltage drop along the I_D current flow (Fig. 1) is

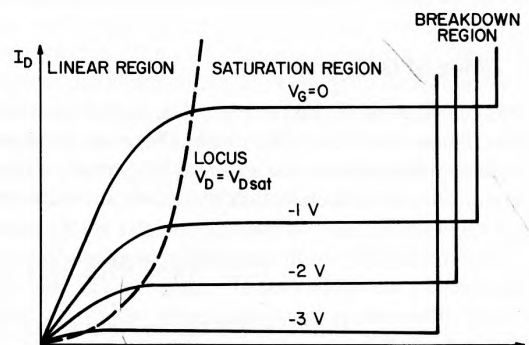


FIGURE 3 n -Channel JFET current-voltage static characteristics.

$$dV = I_D dR = \frac{I_D dy}{2[d - W(y)]Zq\mu_n N_D} \quad (3)$$

where dR is the incremental channel resistance, μ_n the electron mobility, q the magnitude of the electronic charge, N_D the ionized donor density of the n channel, and Z the channel width in the z direction. At any point y along the channel, the depletion region thickness $W(y)$ is given by

$$W(y) = \left\{ \frac{2\epsilon_s}{qN_D} [V_i + V_G + V(y)] \right\}^{1/2} \quad (4)$$

where ϵ_s is the permittivity of the semiconductor, V_i the built-in gate junction potential, V_G the magnitude of the gate-to-source voltage, and $V(y)$ the potential in the channel at y . Substitution of Eq. 4 in Eq. 3 and integration between $V = 0$ at the source ($y = 0$) and V_D at the drain ($y = L$) yields the drain current magnitude:

$$I_D = G_0 \left\{ V_D - \frac{2}{3} \frac{1}{V_p^{1/2}} \times [(V_D + V_i + V_G)^{3/2} - (V_i + V_G)^{3/2}] \right\} \quad (5)$$

with

$$G_0 = 2qdZ\mu_n N_D / L \quad (6)$$

Here G_0 is the undepleted channel conductance, and V_p the magnitude¹ of the gate-to-source pinch-off voltage at which the gate depletion regions meet for $V_D = 0$. From $p-n$ junction theory,

$$V_p = qd^2 N_D / 2\epsilon_s \quad (7)$$

For small V_D values such that $V_D \ll (V_i - V_G)$, Eq. (5) reduces to

$$I_D = G_0 \left[1 - \frac{(V_i + V_G)^{1/2}}{V_p^{1/2}} \right] V_D \quad (8)$$

which shows the linear dependence of the drain current on the drain voltage. That is, the device behaves like a voltage-controlled resistor for small drain voltages. As would be expected from the assumption of an abrupt gate junction, I_D varies as a square root of the gate voltage V_G .

With the drain voltage applied, the gate depletion region thickness is the largest at the drain end of the channel and the smallest at the source end due to the voltage drop along the channel.

At sufficiently high positive drain voltage V_D the gate depletion regions become large enough to start merging at the drain side, the channel is pinched-off and the current flows in part through the depletion region. Beyond this point any further increase in the drain voltage increases

¹ V_p includes the built-in potential V_i .

the drain current only insignificantly: The drain current saturates. Equation (5) for I_D is applicable only up to this point. The drain voltage $V_D = V_{Dsat}$ at which the saturation occurs for long channel devices is given by

$$V_{Dsat} = V_p - (V_i + V_G) \quad (9)$$

the saturation voltage is smaller for larger absolute values of V_G . For $V_D > V_{Dsat}$ the excess drain voltage is dropped across the depletion region.

Substitution of Eq. (9) for V_D in Eq. (5) gives I_{Dsat} , the magnitude of the drain current in saturation as

$$I_{Dsat} = I_p \left[1 - 3 \left(\frac{V_i + V_G}{V_p} \right) + 2 \left(\frac{V_i + V_G}{V_p} \right)^{3/2} \right] \quad (10)$$

where I_p , the pinch-off current, is

$$I_p \equiv q^2 \mu_n Z N_D^2 d^3 / 3\epsilon_s L \quad (11)$$

Therefore, to the first approximation I_{Dsat} is independent of the drain voltage.

For drain voltages in excess of V_{Dsat} the depletion region moves somewhat toward the source, reducing the channel length and thus slightly increasing the drain current so that the current-voltage characteristics in actual devices show a small but finite slope in the current saturation region. The shorter the channel length, the more pronounced is the effect.

In the third region of the current-voltage characteristics the drain current increases extremely rapidly because of the avalanche breakdown of the gate junction that occurs when the drain-to-source voltage is

$$V_D = V_B - V_G \quad (12)$$

where the absolute voltage values are used; V_B is the gate junction breakdown.

For devices with short channels (small $L/2d$ ratios) the equations relating the drain current to the drain and gate voltages become rather inaccurate. One of the reasons is that the electron mobility μ_n is not constant as assumed but is a function of the electric field. When the field reaches the magnitude of 3 to 5×10^4 V cm⁻¹ the electron velocity in silicon crystal reaches its maximum possible saturation value $v_s \approx 10^7$ cm sec⁻¹. In a short channel, electric fields may exceed the critical field value. In this situation, keeping all biases unchanged, the drain current is reduced by a factor of $1 + (\mu_0 V_D / v_s L)$, with μ_0 being the low-field electron mobility.

The current-voltage characteristics of Fig. 3 are described as pentodelike. In JFETs with very short channels that are depleted even in the absence of the external gate potential, the potential barrier at the source prevents the drain current from flowing at low positive drain voltages. A high enough drain voltage lowers the barrier height

and allows current to flow. In such devices and for specific geometries the drain current may not saturate and the current-voltage characteristics are triodelike.

C. Small-Signal Parameters

At the output, the dc drain current I_D is a function of both the dc source-drain voltage V_D and the dc gate voltage V_G . The total drain current, the sum of the small-signal drain current, i_D and the dc drain current I_D , is then a function of the total drain and gate voltages. The small-signal drain current can then be written as,

$$i_D = I_D(V_D + v_D, V_G + v_G) - I_D(V_D, V_G) \quad (13)$$

Expanding the total drain current in Eq. (13) to first-order obtains,

$$I_D(V_D + v_D, V_G + v_G) = I_D(V_D, V_G) + \left. \frac{\partial I_D}{\partial V_D} \right|_{V_G} v_D + \left. \frac{\partial I_D}{\partial V_G} \right|_{V_D} v_G \quad (14)$$

Substituting Eq. (14) into Eq. (13) obtains,

$$i_D = \left. \frac{\partial I_D}{\partial V_D} \right|_{V_G} v_D + \left. \frac{\partial I_D}{\partial V_G} \right|_{V_D} v_G \quad (15)$$

Using the following definitions,

$$g_D \equiv \left. \frac{\partial I_D}{\partial V_D} \right|_{V_G} \quad g_m \equiv \left. \frac{\partial I_D}{\partial V_G} \right|_{V_D} \quad (16)$$

the small-signal drain current becomes,

$$i_D = g_D v_D + g_m v_G \quad (17)$$

The quantity, g_D is called the drain or channel conductance and g_m is called the transconductance. Using these parameters, a simplified small-signal (incremental) equivalent circuit for the common source configuration and valid only for low frequencies can be constructed. This circuit is shown in Fig. 4.

The channel conductance, g_D , in the linear region of the characteristic can be determined by differentiating Eq. 8.

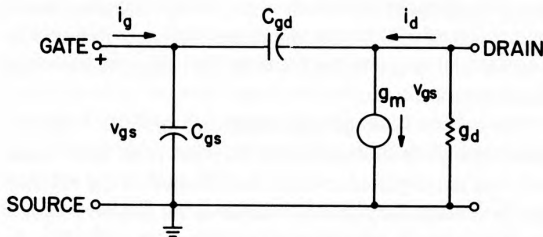


FIGURE 4 JFET incremental low-frequency equivalent circuit.

The channel conductance is given then as

$$g_D \equiv \left. \frac{\partial I_D}{\partial V_D} \right|_{V_G=\text{const}} = G_0 \left[1 - \frac{(V_i + V_G)^{1/2}}{V_p^{1/2}} \right] \quad (18)$$

which is valid for drain voltages $V_D \ll (V_i - V_G)$.

In an ideal JFET the incremental conductance in the saturation region is zero, but in actual devices it has some small value for the reasons mentioned in Section II.B.1.

The effectiveness of the gate voltage control over the drain current is measured by the trans-conductance,

$$g_m \equiv \left. \frac{\partial I_D}{\partial V_G} \right|_{V_D} \quad (19)$$

For the linear region, g_m can be obtained by differentiating Eq. (8) for I_D with respect to V_G keeping V_D constant; g_m reaches its maximum in the saturation region. Differentiating $I_{D\text{sat}}$ [Eq. (10)] with respect to V_G , one finds that $g_{m\text{sat}}$ is equal to g_d ; that is, it is equal to the conductance in the linear region.

In a practical device the conductance g_d is smaller than that given by Eq. (18) because of the existence of small resistances in series with the source and drain. The source resistance also lowers the value of g_m , as given by Eq. (19).

The gate capacitance C_G , although of a distributed nature, usually can be accurately represented by the sum of two capacitances C_{GS} and C_{GD} . The first is the capacitance between the gate and the source, and the second is the capacitance between the gate and the drain as shown in Fig. 4:

$$C_G = C_{GS} + C_{GD} \quad (20)$$

These are intrinsic device capacitances. At high frequencies it is necessary to take extrinsic parasitic capacitances into account as well. In very short channel transistors the parasitic capacitive and inductive parameters may be most important.

The cutoff frequency f_T is the frequency at which the input current becomes equal to the output current with the output short-circuited (Fig. 4). It follows from this definition that

$$f_T = g_m / 2\pi C_G \quad (21)$$

It can be shown that in saturation the maximum possible value for f_T is

$$f_{T\text{max}} = \frac{q\mu_n N_D d^2}{4\pi\epsilon_s L^2} \quad (22)$$

The cutoff frequency is obviously higher for small channel lengths and for n -channel devices in silicon because μ_n , the electron mobility, is much higher than the hole mobility μ_p .

In devices with channels shorter than $2\mu\text{m}$ and operating in the saturated region of the current-voltage characteristics, electrons mainly flow in that channel

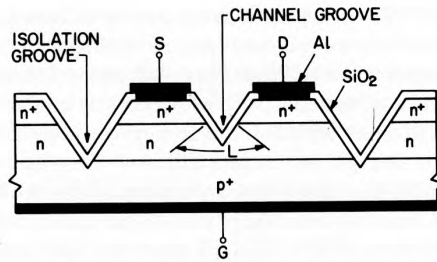


FIGURE 5 Schematic representation of the V-groove channel JFET. [After Mok, T. D., and Salama, C. A. T. (1978). *IEEE Transactions on Electron Devices*, ED-25, No. 3. Copyright 1978. IEEE.]

region, where the electric field is high and the electron drift velocity reaches its saturation value $v_s = 10^7$ cm sec⁻¹. As a result, the saturated drain current, I_{Dsat} , the transconductance, $g_{m sat}$, and the cutoff frequency, f_T , all become directly proportional to the saturated electron velocity v_s .

D. Silicon JFET Structures

A lateral JFET can be fabricated by double-diffusion techniques, making the process compatible with bipolar transistors integrated on the same silicon chip. The gate and channel regions are diffused or implanted from the top surface into an epitaxial silicon layer deposited on a silicon substrate.

A structure known as a V-groove channel JFET (Fig. 5) is obtained by preferential (anisotropic) etching of a groove in the (100)-oriented n -type epitaxial silicon layer deposited on the p^+ substrate. This $p^+ - n$ junction acts as a single gate. The top n^+ layers are diffused or ion-implanted to achieve good ohmic contact to subsequent metallization. In addition, etched isolation grooves are necessary to achieve electrical isolation between individual JFETs or other devices on the same silicon chip.

As the reverse bias on the p^+n junction increases, the depletion region moves up, cutting into the electron current path. Small channel lengths and high packing density on the silicon chip are achievable with this technique.

Still another JFET structure, the static induction transistor (SIT), is analogous to a vacuum tube (Fig. 6). p^+ -Doped regions (gates) are embedded in the n -type silicon, forming a grid interposed between the source and drain electrodes. Depletion regions formed between the gates modulate the electron current flow from the source to drain. This multichannel device can be constructed to exhibit triodelike characteristics.

E. Metal Schottky-Barrier Field-Effect Transistor (MESFET)

The JFET $p^+ - n$ junction gate can be replaced by a Schottky barrier formed by metallizing the n -type semiconduc-

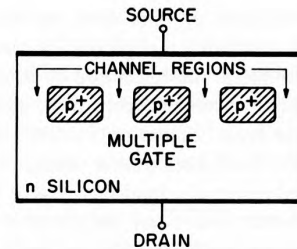


FIGURE 6 Static induction transistor. (Schematic representation.)

tor surface. MESFETs can be fabricated in silicon and other semiconductors. Here, we describe a GaAs device that is particularly useful in very high speed computer circuits and in microwave frequencies above 4 GHz.

An n -type GaAs layer is deposited on a semi-insulating GaAs substrate (doped, e.g., with chromium) to make it electrically nonconducting (Fig. 7a). The substrate insulates the active device electrically, but not thermally, from the heat sink. GaAs is well suited for microwave operation because its low-field electron mobility is five times higher than that of silicon, and its peak electron drift velocity is twice as large as silicon. At fields lower than 3×10^3 V cm⁻¹ the drift velocity reaches its peak value $v_p = 2 \times 10^7$ cm sec⁻¹ and then decays to 50% of this value as the field increases to $\sim 2 \times 10^4$ V cm⁻¹. As Fig. 7b indicates the drift velocity peaks at y_1 , close to the channel center, then drops to $v_s = 10^7$ cm sec⁻¹ at y_2 under the drain edge of the gate, where the field is the highest. Since the electrons' velocity decreases, there is an accumulation of electrons to preserve current continuity. As the channel thickness starts increasing, the field decreases and the electron velocity increases, eventually reaching

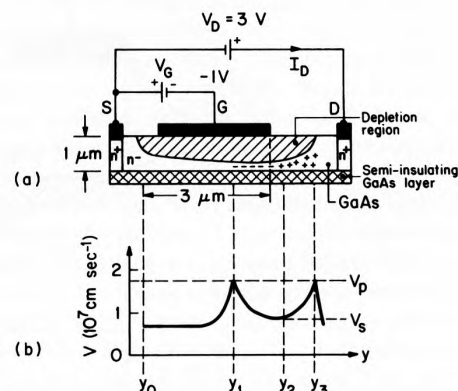


FIGURE 7 (a) Schematic cross section of a MESFET; (b) electron velocity as a function of positive y . [After Liechti, A. (1976). *IEEE Transactions on Microwave Theory and Techniques*, MTT-24, No. 6. Copyright 1976, IEEE.]

the saturation velocity, v_s . This time, in order to maintain the current continuity a depletion of electrons takes place. This amounts to the formation of positive charge. Thus a dipole layer forms underneath the gate edge. The presence of this layer affects unfavorably such device properties as the drain breakdown voltage and forward voltage drop and power dissipation.

For a gate shorter than $1\text{ }\mu\text{m}$ the electron transit time from source to drain may be shorter than the electric relaxation time of the semiconductor. This results in an electron velocity that exceeds several times the steady-state peak velocity v_p and in a considerably improved device frequency response. In these devices only the extraneous parasitic reactances and resistances restrict the high-frequency performance.

F. Modulation-Doped Field-Effect Transistor

The modulation-doped field-effect transistor (MODFET) is a modified MESFET device that seeks to optimize speed and current-carrying capability. In a MESFET the dopants that provide the necessary charge carriers and the source-drain current flow both reside in the top, thin semiconductor layer. The close proximity of the ionized dopant atoms and the charge carriers leads to a strong Coulomb interaction that in turn produces strong impurity scattering. As a consequence of the high impurity scattering rate, the carrier mobility is reduced resulting in relatively slow device performance. Though reducing the doping concentration would have the beneficial result of enhancing the mobility, it would also reduce the conductivity and hence the current carrying capacity of the device. Therefore, a trade-off exists in a standard MESFET structure.

MODFETs utilize modulation doping to increase the free carrier concentration without a concomitant decrease in the carrier mobility. When a highly doped wide band gap semiconductor material, such as AlGaAs, is used to form a heterostructure with an intrinsic narrower gap semiconductor, such as GaAs, carriers will move from the highly doped layer into the intrinsic layer to equilibrate the Fermi level. In this way, free charge carriers are introduced into the GaAs layer without adding dopant impurities. The spatial separation of the charge carriers in the GaAs from the ionized dopants in the AlGaAs results in a reduced Coulomb interaction and consequently a much lower ionized impurity scattering rate and higher mobility.

A typical MODFET (Fig. 8) consists of an undoped GaAs layer, $1\text{ }\mu\text{m}$ thick or less, deposited on the top of a semi-insulating GaAs layer. A very thin (a few nanometers thick) undoped, insulating AlGaAs film is grown by molecular beam epitaxy on top of the undoped GaAs layer. Then a thicker ($\sim 50\text{-nm}$) n -type doped AlGaAs layer is epitaxially grown on the undoped layer. A metal gate is finally deposited on this film, forming a Schottky barrier. The source and drain are both electrically connected to the opposite sides of the n AlGaAs film and the two-dimensional gas region.

The two-dimensional electron gas concentration can be changed by varying the gate voltage; this in turn modulates the magnitude of the drain current. In the MESFET, on the other hand, the thickness of the channel is modulated and not the charge concentration.

The purpose of the undoped intrinsic AlGaAs layer is to avoid Coulomb scattering of electrons at the heterojunction interface and thus to increase the electron mobility. Depletion- or enhancement-type FETs can be

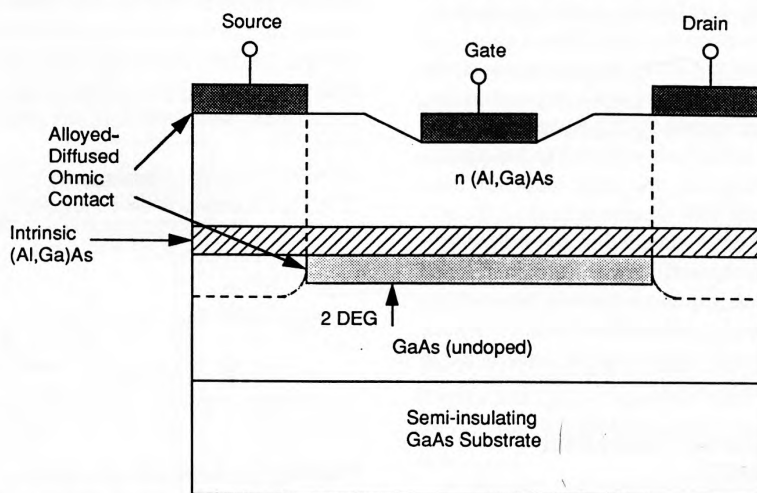


FIGURE 8 Schematic (not to scale) cross section of MODFET [After T. J. Drummond, W. T. Masselink, and H. Morkoç (1986). *Proceedings of the IEEE*, Vol. 74, No. 6, Copyright 1986, IEEE].

made depending on the appropriate choice of the undoped AlGaAs layer thickness and gate polarity. In the case of the depletion-type device the unbiased Schottky-barrier depletion region reaches only the heterojunction interface. For the enhancement-type device the depletion region also encompasses the two-dimensional electron gas.

The MODFET with some structural variations is also known as a high electron mobility transistor (HEMT), two-dimensional electron gas FET (TEGFET), or selectively doped heterojunction transistor (SDHT).

GaAs and AlGaAs have almost perfectly matched crystal lattices, so there is no strain introduced at their interface. In some other compound semiconductor structures such as InGaAs/GaAs used to experimentally fabricate MODFETs, lattices are considerably mismatched. Such MODFETs are known as pseudomorphic (pseudo-isomorphic) or strained quantum well transistors (SQWFETs). A 10–20 nm thick, *narrow* energy bandgap InGaAs layer is sandwiched between two layers of relatively thick (200–1000 nm) *wider* energy bandgap GaAs. The atomic spacing in the very thin film of InGaAs adjusts itself to the atomic spacing of GaAs; thus, the creation of crystal dislocations and deleterious interface states is avoided.

The saturated electron velocity in the InGaAs crystal, at liquid nitrogen temperature (77 K) is 20% higher than that in GaAs. This results in MODFETs with superior, high-frequency response and higher switching speeds than are possible with the AlGaAs/GaAs system. These devices are known as pseudomorphic HEMTs or PHEMTs.

Normally, electrical conduction in the valence band takes place by the transport of slow “heavy” holes. The introduction of the compressive strain into the InGaAs crystal modifies the valence band structure so that fast “light” holes with small effective masses become available for electrical conduction. This makes the fabrication of complementary *p*-type MODFETs with speeds comparable to their *n*-type counter-parts possible.

G. JFET Applications

Silicon JFETs are particularly useful in those applications that require high input impedance and low leakage currents and often are integrated with bipolar transistor circuits on the same silicon chip. Process compatibility between the two types of devices is often easier to achieve with the *p*-channel devices. Junction field-effect transistors operated in the linear region of characteristics are frequently used as voltage-controlled resistors.

By using capacitors combined with JFETs operating in the linear region of characteristics, it is simple to fabricate voltage-controlled phase shifters. Other applications include voltage-controlled attenuators, variable-gate

amplifiers, voltage tuned-filter circuits, and operational amplifiers.

Junction field-effect transistor switches show excellent on-to-off current ratios and are capable of switching either very small currents, of the order of microamperes, or very large currents of the order of tens of amperes, as in field-effect power transistors.

GaAs- and InP-based MESFETs and MODFETs are primarily utilized in high-frequency device applications that arise mainly in wireless telecommunications. The key drivers in wireless RF device design are efficiency, power, and frequency performance at high linearity. Of the competing device technologies, GaAs PHEMTs offer the highest f_t at the lowest noise figure. Commercial MODFET structures used in power amplifiers have been reported with operating frequencies up to 40 GHz. Most commercial PHEMTs reported operate between 30 and 40 GHz, but some have been reported with operating frequencies up to 100 GHz.

IV. METAL-INSULATOR-SEMICONDUCTOR FIELD-EFFECT TRANSISTOR

The control electrode of a metal-insulator-semiconductor field-effect transistor MISFET is separated from the channel region by an insulator. The most important MISFET is the metal oxide semiconductor transistor (MOSFET). Its most widely used version utilizes silicon as the semiconductor and silicon dioxide (SiO_2) as the gate insulator. Virtually all semiconductor microprocessors and memory chips are fabricated using silicon-based MOSFETs. They are relatively easy to fabricate, can be readily miniaturized and integrated, and are highly reliable and reproducible. The combination of these meritorious features has conspired to make silicon MOSFETs by far the most attractive devices for computing applications. At this writing, commercially available integrated circuits employing MOSFETs with 0.13- μm gate lengths are becoming available. At these and smaller device dimensions gigabit scale integration will become commercially possible.

The *n*-channel MOSFET (Fig. 9) consists of the *p*-type bulk (substrate) and two n^+ regions: the source *S* and the drain *D*. The metal gate is insulated from the bulk by a thin SiO_2 layer. The source is kept at ground potential for the purpose of this discussion, and the drain is positively biased with respect to the source (V_D). In the absence of the positive gate potential V_G the source and drain pn^+ junctions allow only leakage current to flow. For a sufficiently high positive V_G , an *n*-type inversion layer (channel) is formed electrically, connecting the source to the drain and permitting electrons to flow. This channel current can be

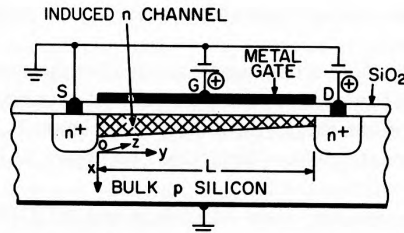


FIGURE 9 Schematic representation of an *n*-channel MOSFET.

controlled by varying the gate voltage. The device is said to be operating in the enhancement mode.

By changing the *n*- and *p*-region doping types and making V_G and V_D negative, one obtains a *p*-channel MOSFET.

The depletion mode MOSFET is provided with a conducting channel during its fabrication by an ion implantation or diffusion process. Such a MOSFET is on (conducts current) in the absence of gate bias. An application of a negative gate voltage V_G will modulate the *n*-channel device current flow. Similarly, positive V_G is used for the *p*-channel depletion mode MOSFET. The symbols for the four possible MOSFET versions are shown in Fig. 10.

To understand the MOSFET operation it is essential to understand the behavior of the silicon surface. This can be described with the aid of a metal oxide semiconductor (MOS) capacitor structure.

A. MOS Capacitor

The MOS capacitor described (see also Fig. 11a) consists of the metal electrode (gate) insulated from the *p*-type silicon substrate by a thin SiO_2 layer. If the metal's work function $q\psi_m$ (q is the electron charge magnitude) is smaller than the work function $q\psi_{Si}$ of silicon, then as soon as the two materials are brought to the opposing surfaces of the oxide layer, some electrons are transferred from the gate

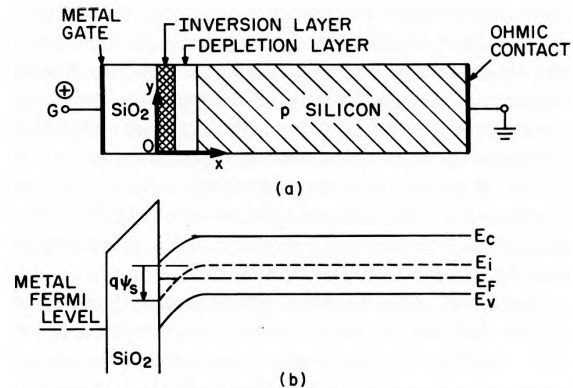


FIGURE 11 MOS capacitor. (a) Schematic cross section; (b) energy band diagram for strong inversion.

to the substrate by the imperfect insulation external to the capacitor. Thus the gate becomes positively charged and therefore repels holes in the silicon at the Si-SiO₂ interface. The *p*-type silicon at the surface becomes negatively charged as the number of holes is now smaller than expected from the silicon doping density. Consequently, at the surface the silicon energy bands bend so that the silicon intrinsic energy level E_i comes closer to the silicon Fermi level; the Fermi levels of the metal and silicon align, and the device comes to equilibrium (not shown in the figure). The total band bending is measured by the ψ_s (Fig. 11b), sometimes called surface potential; ψ_s is positive when pointing down. The application of an external potential to the gate equal to the potential difference $\psi_m - \psi_{Si} = \psi_{ms}$ (as modified by the presence of the SiO_2 layer) will return the bands to their initial flat-band condition. The gate voltage $V_{FB1} = \psi_{ms}$ required to do so is known as the flat-band voltage. It is dependent on the kind of metal used and the polarity and density of the semiconductor doping. For an aluminum gate and *p*-type silicon doped to 10^{15} acceptors cm^{-3} , $V_{FB1} \simeq -0.9$ V.

As mentioned in Section II, depending upon the external gate voltage, the MOS structure can be biased into accumulation, depletion or inversion. The metal oxide-silicon system encountered in practice typically contains several sources of impurity charge located at various sites. These charges can affect the bias condition of the semiconductor layer. For example, impurity charge in the oxide can produce inversion even in the absence of an external bias. At ~ 2 nm from the Si-SiO₂ interface a fixed positive charge exists in the oxide. This charge, Q_{ss} , results from the imperfect oxidation of silicon ions. It is a function of crystal orientation and is smallest for the (100) crystal orientation. By the use of very careful processing and annealing, the number of charges (Q_{ss}/q) can be reduced to less than

MODE OF OPERATION MOSFET TYPE	ENHANCEMENT (NORMALLY OFF)	DEPLETION (NORMALLY ON)
<i>n</i> -TYPE CHANNEL		
<i>p</i> -TYPE CHANNEL		

FIGURE 10 MOSFET symbols.

10^{10} cm^{-2} . Other charges, such as mobile positive ions in SiO_2 due to the presence of sodium or potassium ions, and charged surface states, are kept to a minimum by various techniques and have been omitted from this discussion for the sake of simplicity.

The capacitance C_{ox} per unit area of the capacitor formed by the gate and the semiconductor in a highly conductive state is given by

$$C_{ox} = \epsilon_s / d_{ox} \quad (23)$$

where ϵ_s is the semiconductor permittivity and d_{ox} the oxide thickness. The total flat-band voltage V_{FB} that accounts both for the work-function difference and Q_{SS} -induced potential is therefore given by

$$V_{FB} = \psi_{ms} - (Q_{SS} / C_{ox}) \quad (24)$$

For $V_G > V_{FB}$, holes from the p -type silicon surface are repelled, and a surface depletion region consisting of negative acceptors is formed. Hence, the energy bands bend downward (Fig. 11b). Further increase of the gate potential results in the attraction of electrons to the silicon surface, inverting it from p - to n -type. In this condition the energy bands bend enough to make the intrinsic energy level E_i cross the Fermi level (Fig. 11b). The density of electrons n induced at the surface by the gate voltage can be calculated from

$$n = \frac{n_i^2}{N_A} \exp \frac{q\psi_s}{kT} \quad (25)$$

where n_i is the silicon intrinsic concentration ($n_i^2 \simeq 2 \times 10^{20} \text{ cm}^{-3}$), and N_A —the density of the fully ionized acceptors in the substrate.

Strong inversion takes place when the electron concentration in the inversion layer of the surface is equal to the hole density in the p bulk. This takes place for band bending $\psi_s = 2\psi_F = 2E_F/q$ (E_F is Fermi energy level). The n -type inversion layer forms an abrupt, practically one sided junction with the p -type bulk. The depletion region thickness of this junction is

$$W_{Dmax} = \left(\frac{4\epsilon_s \psi_F}{q N_A} \right)^{1/2} \quad (26)$$

For $\psi_s > 2\psi_F$, W_{Dmax} remains almost unchanged, since the inversion layer acts as an electrostatic shield preventing the electric field from penetrating the bulk further.

A sufficiently high negative gate voltage repels electrons and attracts holes to the silicon surface, creating an accumulation layer of holes and making the silicon surface more p -type. The energy bands bend upward for this surface condition, and the Fermi level comes closer to the valence band: the band bending surface potential ψ_s becomes *negative*. The concentration of holes at the surface of the accumulation layer is given by

$$p = N_A \exp(-q\psi_s/kT) \quad (27)$$

1. Threshold Voltage

The turn-on gate voltage for strong inversion is known as the threshold voltage V_T . It must be at least equal to the sum of the flat-band voltage V_{FB} , the band bending surface potential ψ_s , and the voltage dropped across the depletion region,

$$V_T = V_{FB} + 2\psi_F - (Q_{Dmax} / C_{ox}) \quad (28)$$

where

$$Q_{Dmax} = -q N_A W_{Dmax} \quad (29)$$

is the total charge per unit area in the depletion region of thickness W_{Dmax} . Equation (28) still applies in the case where, for ease in fabricating small structures, the metal gate is replaced by a heavily doped p^+ or n^+ polycrystalline silicon layer. Appropriate changes in the value of V_{FB} reflect the change in gate material.

B. Current-voltage Relations in the MOSFET

In the MOSFET, the current flowing in the channel under strong inversion (Fig. 9) causes a resistive voltage drop so that the band bending ψ_s varies with the distance y from the origin at the source as

$$\psi_s(y) = 2\psi_F + V(y) \quad (30)$$

Here, $V(y)$ is the reverse bias between the elemental section of the channel and bulk at y (Fig. 9). In view of Eq. (30) the depletion region thickness as a function of position y is given by

$$W'_{Dmax} = \left\{ \frac{2\epsilon_s [2\psi_F + V(y)]}{q N_A} \right\}^{1/2} \quad (31)$$

Accordingly, the depletion region charge per unit area is given by

$$Q'_{Dmax}(y) = -\{2\epsilon_s q N_A [2\psi_F + V(y)]\}^{1/2} \quad (32)$$

and the threshold voltage [Eq. (32)] becomes

$$V'_T = V_{FB} + 2\psi_F + V(y) - (Q'_{Dmax} / C_{ox}) \quad (33)$$

The total charge per unit area $Q_s(y)$ at the silicon surface is the sum of the channel charge per unit area $Q_{ch}(y)$ and the depletion layer charge Q'_{Dmax} per unit area. Hence,

$$Q_{ch}(y) = Q_s(y) - Q'_{Dmax}(y) \quad (34)$$

(For an n -channel MOSFET all these charges are negative.)

The total charge per unit area, Q_s , is equal and opposite in sign to the charge per unit area on the gate, Q_G , that is,

$$Q_s(y) = -Q_G(y) = -C_{ox} V_{ox} \quad (35)$$

Since,

$$V_{ox} = V_G - V_{FB} - 2\phi_F - V(y) \quad (36)$$

$$Q_{ch} = -C_{ox}[V_G - V_{FB} - 2\phi_F - V(y)] - Q'_{Dmax} \quad (37)$$

Therefore, by Ohm's law the drain current is given by

$$I_D = -Z\mu_n Q_{ch}(y) \frac{dV(y)}{dy} \quad (38)$$

where Z is the channel width in the z -direction (Fig. 9). The electron mobility μ_n is assumed here to be constant and not affected by the electric field ($-dV/dy$) magnitude. Integration of Eq. (38) with limits 0 and V_D for $V(y)$, and 0 and L for y , respectively, gives the magnitude of the drain current as

$$I_D = \frac{Z\mu_n C_{ox}}{L} \left\{ \left(V_G - V_{FB} - 2\psi_F - \frac{V_D}{2} \right) V_D - \frac{2}{3} \frac{(2\epsilon_s q N_A)^{1/2}}{C_{ox}} [(V_D + 2\psi_F)^{3/2} - (2\psi_F)^{3/2}] \right\} \quad (39)$$

More exact theories include the effect of the channel region nonuniform doping (if any) and of current carrier diffusion in the channel resulting from the carrier concentration gradients. The concentration is the highest at the source.

1. Linear Region of Characteristics

For small drain voltages such that $V_D < 2\psi_F$ and $V_D < (V_G - V_{FB} - 2\psi_F)$, Eq. (39) reduces to

$$I_D = G_0(V_G - V_T)V_D \quad (40)$$

where $G_0 = Z\mu_n C_{ox}/L$ and V_T is the equilibrium threshold voltage [Eq. (28)]. Analogous equations can be derived for the p -channel MOSFET. Equation (40) demonstrates the linearity of the current-voltage static characteristics for low drain voltages (Fig. 12). The linear region of the current-voltage characteristics is also known as the triode region.

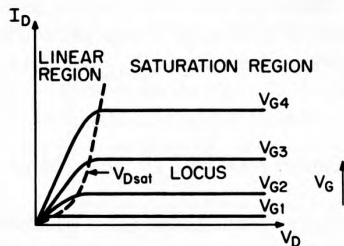


FIGURE 12 Current-voltage characteristics of an n -channel MOSFET.

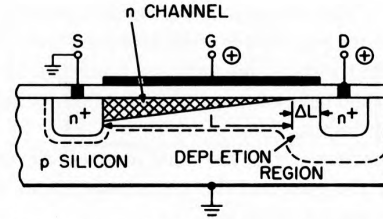


FIGURE 13 n -Channel MOSFET operating in saturation: ΔL represents the thickness of the drain depletion region.

2. Saturation Region of Characteristics

Equation (39) for the drain current shows that I_D first increases linearly with the drain voltage V_D . Then its rate of increase slows down. To a first approximation, at $V_D = V_{Dsat}$ the current reaches its saturated value I_{Dsat} (Fig. 12). Beyond this point, Eq. (39) for I_D loses its validity. At V_{Dsat} , no electrons are induced at the channel drain end ($y = L$): The channel is pinched off, and there is no more gate-induced charge for current conduction. This happens because the drain voltage neutralizes the gate voltage.² Further electron flow occurs in the drain depletion region (Fig. 13), whose length ΔL increases as V_D increases above V_{Dsat} .

The channel charge, $Q_{ch}(y)$ vanishes near the drain, at $y = L$. Therefore, Q_{ch} is equal to zero in Eq. (37). Calling $V(y = L) = V_{Dsat}$, Eq. 37 can be solved for V_{Dsat} to give,

$$V_{Dsat} = V_G - \left(V_{FB} + 2\phi_F - \frac{Q'_{Dmax}}{C_{ox}} \right) \quad (41)$$

Equation (41) can be rewritten using the expression for V'_T given by Eq. (33) recognizing that $V(y)$ is equal to zero under pinch-off conditions. Q_{Dmax} is then equal to Q'_{Dmax} and $V'_T = V_T$. Equation (41) becomes

$$V_{Dsat} = V_G - V_T \quad (42)$$

In silicon devices with gates that are shorter than $2 \mu\text{m}$, the longitudinal electric field in the channel, in many cases, exceeds the critical field at which the electron velocity reaches its saturation value v_s and causes the drain current to saturate under the gate before the channel pinch-off takes place. In this situation Eqs. (41, 42) lose their validity.

The magnitude of the drain current in saturation I_{Dsat} can be obtained from Eq. (39) by substituting Eq. (41) for V_{Dsat} . For moderately or lightly doped substrates and for oxides that are thin compared with the surface depletion

²In actuality there is, of course, a small remaining charge at the edge of the depletion region in order to make the current flow possible.

region thickness, I_{Dsat} for both p - and n -channel devices can be approximated by the quadratic relationship³

$$I_{Dsat} = \frac{1}{2} G_0 (V_G - V_T)^2 \quad (43)$$

Equation (43) demonstrates the saturation drain current independence of the drain voltage for an idealized device. In actual devices the drain conductance dI_{Dsat}/dV_{Dsat} is not equal to zero. As the drain voltage is increased above V_{Dsat} the pinch-off point moves slightly toward the source because of the increased drain depletion region thickness (Fig. 13). This reduces the channel length slightly, thus increasing the current. This effect is much more pronounced in short-channel devices where the channel length is of the order of few micrometers or less.

Another complication arises for drain voltages higher than V_{Dsat} because the drain field starts dominating the gate field at the channel end, and the current spreads away from the surface so that the simple theory based on the gradual channel approximation loses some of its validity.

3. Subthreshold Behavior

When the gate voltage is below the threshold voltage in an enhancement mode device, the semiconductor layer is no longer in strong inversion. This region of operation of the device is called cutoff. In cutoff, the relationships derived above for the drain current are no longer valid. Though the drain current under these conditions is not very large, it is, nevertheless, not necessarily equal to zero either. The current that flows between the source and the drain below threshold is called the subthreshold current. The subthreshold current is extremely important in digital circuit design. In digital circuits a MOSFET is switched into cutoff in order to turn the drain current off. Ideally, no drain current should flow in cutoff. In general a small subthreshold current does flow. A nonzero subthreshold current within the device still draws power even when it is switched off. This results in unnecessary power consumption reducing battery lifetimes in portable electronic applications and heating of the chip. The subthreshold current is dominated by diffusion due to the relatively large channel resistance in cutoff.

4. Body Bias Effect

In some circuits using MOSFETs it is necessary to introduce a reverse bias V_{BS} between the body and source. This voltage appearing between the channel and body increases the channel-body depletion region thickness and therefore the total depleted charge Q_{Dmax} . As a result, the

³More accurate analysis shows that for heavily doped substrates the one-half factor is replaced by a factor smaller than one-half and dependent on the substrate concentration.

threshold voltage V_T is increased by an increment ΔV_T that is roughly proportional to $|V_{BS}|^{1/2}$ for body biases of a few volts.

C. Short-Channel MOSFET Operation

The quantitative expressions derived above for the current-voltage characteristic for a MOSFET are based on the long-channel approximation. In this approximation, it is assumed that the longitudinal electric field is negligible with respect to the transverse field. However, as the channel length decreases, the source and drain-depletion layer thicknesses become comparable with the channel length and encroach on the area under the gate. As a result, the field becomes effectively two-dimensional. For current state-of-the-art devices, the feature lengths are sufficiently small that the long-channel approximation is no longer valid. Reduction of the gate length to ultrasubmicron dimensions influences device behavior in several important ways.

There are many different effects that appear at short-channel lengths that greatly alter device performance. These effects arise from three primary sources: (1) the electric-field profile becomes two-dimensional; (2) the electric field strength in the channel becomes very high; and (3) a decrease in the physical separation between the source and drain. The two-dimensionality of the potential profile causes a reduction in the threshold voltage of the device as well as a reduction of the mobility by gate-induced surface fields. The presence of a high electric field strength in the channel leads to velocity saturation, impact ionization near the drain, gate oxide charging, and parasitic bipolar effects. The decrease in the physical separation of the source and drain results in punchthrough and channel-length modulation. The effect of some of these mechanisms on device behavior can be summarized as follows.

1. Threshold Voltage Variation

As the channel length decreases a smaller surface depletion region charge is controlled by the gate potential (Fig. 14a). This results in a smaller value of the threshold voltage, V_T , from that of a comparable long-channel device. A channel with an extremely small width, Z , however, may show a net increase in the threshold voltage because of the lateral spreading of the depletion region in the z -direction due to the presence of fringing fields (Fig. 14b).

The threshold voltage can be intentionally altered by implanting p - or n -type dopants through the gate-oxide layer. Very precise quantities of impurities can be added in this manner and the threshold voltage can be carefully altered.

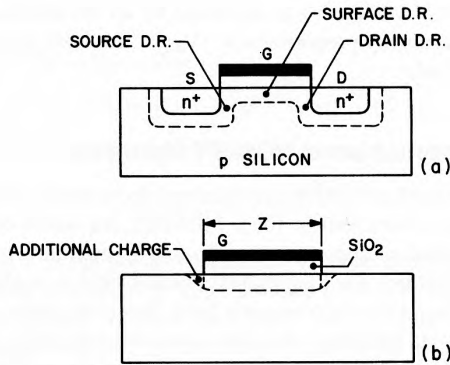


FIGURE 14 (a) Overlapping of the source and drain depletion regions (D.R.) with the surface depletion region; (b) cross section along the z axis showing additional charges induced by the fringe fields.

2. Surface Drift Mobility

In the expressions for the current-voltage characteristic it is assumed that the carrier mobility is constant, independent of the channel depth in the x -direction and of the electric field strength. The mobility at the Si-SiO₂ interface, often called the surface mobility, is typically smaller than that within bulk silicon, due to the presence of interface roughness scattering and surface charge scattering that reduce the mobility. However, at small gate lengths, the carrier mobility is further lowered by the action of the transverse fields. The high gate field more strongly attracts the charge carriers to the imperfect surface reducing the mobility.

3. Saturation Drift Velocity

If the voltage is not appropriately scaled as the channel length decreases, the electric field within the channel increases substantially. In silicon, the electron and hole drift velocities saturate at applied electric field strengths in excess of ~ 100 kV/cm. In devices with very short channel lengths, the electrons move at the saturated velocity over most of their current path. Hence, the saturation of the drain current may occur at a significantly lower voltage before the expected drain pinch-off voltage is reached. The velocity saturated current in the limit as the channel length, L , goes to zero is

$$I_{Dsat} = ZC_{ox}(V_G - V_T)\nu_s \quad (44)$$

4. Gate-Oxide Charging

The very high source-drain electric field heats the carriers to very high kinetic energy near the drain. In an n -channel device, the electron charge carriers can reach sufficient

kinetic energy such that they can transfer from the silicon into the oxide. The actual amount of energy that the electrons need to absorb is quite high. Nevertheless, quantum-mechanical tunneling and barrier-lowering effects can reduce the potential barrier of the oxide. Carrier injection leads to gate oxide charging which in turn alters the threshold voltage of the device. The injected charge also makes V_T unstable, degrades the transconductance, and causes excessive gate current. The threshold voltage shift occurs over time. With continued exposure to high electric fields, the injected electron concentration within the oxide increases, gradually changing the threshold voltage. Therefore, gate-oxide charging is a long-term degradation mechanism that is particularly difficult to compensate.

5. Drain-Source Voltage Breakdown and Parasitic Bipolar Effect

The MOSFET drain-to-source breakdown voltage for short gate devices is considerably different from the avalanche breakdown voltage V_B of the drain-bulk junction because of the presence of an $n^+ - p - n^+$ parasitic transistor in parallel with the n -channel MOSFET. As the source current flows into the avalanching drain region, a strong carrier multiplication takes place. The generated electrons flow to the positively biased drain electrode, whereas holes go to the substrate, ground, and also directly to the source. The voltage drop caused by this hole current in the bulk forward biases the source junction, which becomes the emitter of the parasitic bipolar transistor. The drain-to-source breakdown voltage becomes that of an npn transistor with open base, BV_{CEO} . This voltage is dependent on the drain-bulk junction breakdown, the gate length, and the electron diffusion length, in the substrate. It is lower than V_B for short gate lengths and long diffusion lengths.

The existence of the parasitic $n - p - n$ transistor makes the MOSFET susceptible to the destructive thermal instability known as second breakdown. During an inductive circuit turn-off, the drain voltage may become very high while a large drain current is flowing. The high voltage causes the drain to go into avalanche breakdown. The flow of avalanche-generated holes toward the substrate and source develops a very high emitter (source) bias, establishing a positive feedback mechanism that may lead to device destruction.

6. Punchthrough

As the channel length is decreased, the physical separation of the source and drain regions necessarily decreases. Punchthrough arises when the channel length is very small. Under this condition, the source and drain-depletion regions can touch, resulting in a large increase in the drain current.

D. Incremental (Dynamic) MOSFET Parameters

As in the case for the JFET the small signal drain current can be expressed in terms of the channel conductance and the transconductance using Eq. (17).

The incremental channel conductance in the linear region, $\partial I_D / \partial V_D|_{V_G} = g_D$, is obtained by differentiating Eq. (40), which yields

$$g_D = G_0(V_G - V_T) \quad (45)$$

where g_D is linearly dependent on the gate voltage. Its reciprocal is the channel resistance in the linear portion of the current-voltage characteristics (Fig. 12).

The transconductance in the linear region is obtained by differentiating Eq. (40) with respect to V_G with $V_D = \text{const}$

$$g_m = G_0 V_D \quad (46)$$

g_m reaches its maximum value in the saturation region. The differentiation of $I_{D\text{sat}}$ [Eq. (43)] with respect to V_G (keeping V_D constant) results in

$$g_{m\text{sat}} = G_0(V_G - V_T) = G_0 V_{D\text{sat}} \quad (47a)$$

If the saturation current $I_{D\text{sat}}$ is limited by v_s (Eq. 47a) then the transconductance becomes equal to (44)

$$g_{m\text{sat}} = ZC_{ox} v_s \quad (47b)$$

Equations (44)–(47) are valid for idealized devices.

1. Incremental Equivalent Circuit and Cutoff Frequency

The simplified small-signal equivalent circuit of a MOSFET is essentially the same as that of a JFET (Fig. 4). All the extraneous parasitic capacitances and resistances are omitted, and only the intrinsic device is considered. Because of these omissions the equivalent circuit is valid only at low frequencies.

As in the JFET case the cutoff frequency f_T is given by

$$f_T = g_{m\text{sat}} / 2\pi C_G \quad (48)$$

For the MOSFET in saturation, the drain-to-gate capacitance C_{GD} is very small because of the presence of the wide drain junction depletion region and can be neglected as compared with C_{GS} . Consequently the input capacitance C_G is expressed as

$$C_G \simeq C_{GS} \simeq LZC_{ox} \quad (49)$$

Combining Eqs. (46)–(48), we get

$$2\pi f_T = \mu_n V_{D\text{sat}} / L^2 = 1/T \quad (50)$$

The angular cutoff frequency $\omega_T = 2\pi f_T$ is the reciprocal of the charge carrier transit time in the channel. Hence, one can see the importance of short channel lengths

for high speed or high frequency of operation. In well-designed high-speed devices the parasitic elements should be reduced to a minimum since they determine the high-frequency response.

For very short channels (of the order of $1 \mu\text{m}$) the charge transit may take place entirely in the drain region, where the electron velocity reaches its maximum saturated value v_s . In those cases the angular frequency cutoff is given by

$$\omega_T = v_s / L \quad (51)$$

E. MOSFET Applications

MOSFETs are used in numerous discrete and integrated analog circuits as small-signal and power amplifiers, operational amplifiers, phased-locked loops, and so on. Their most important application, however, is in the digital logic and memory circuits of VLSI devices, where they operate as binary elements. Power dissipation can limit the number of MOSFET devices that are practical in a given circuit. The use of complementary MOS (CMOS) devices in digital circuits represents a way of achieving very low power dissipation. The approach uses n -channel MOSFET in series with a p -channel MOSFET (Fig. 15) integrated on the same semiconductor chip. The high input voltage pulse $V_{in} = V_{DD}$ turns the n -channel transistor T_1 on and the p -channel transistor T_2 off. The low input voltage $V_{in} = 0$ turns T_1 off and T_2 on. The two transistors are never on or off simultaneously. Only a very small leakage current, the subthreshold current, may flow in steady state. Thus, in this quiescent state, no appreciable power is dissipated. The only power dissipation that occurs is that which takes place during the switching transients. This can become significant as the switching frequency increases.

1. MOSFET in Very Large Scale and Ultra-Large-Scale Integration

Continuous miniaturization of MOSFETs is necessary to maintain progress in improving integrated circuit

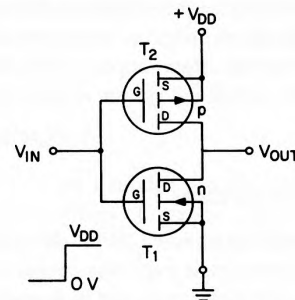


FIGURE 15 Basic CMOS circuit.

performance at lower cost. As shown in the preceding sections simple shrinking of the channel length (below $2\text{ }\mu\text{m}$) has a number of undesirable effects on the MOSFET performance, such as lowering of the threshold voltage; degradation of the drain-to-source breakdown voltage; the possibility of punch-through between the drain and source; and mobile carrier injection into the oxide, leading to device instability. Moreover, the subthreshold current, which adds to the device leakage, increases as the device becomes shorter. Ultra-large-scale integration (ULSI), that approaching gigabit capacity, requires devices with submicron length channels. In the design of such devices it is important to have a simple method for reducing all the device dimensions without detrimental effects.

The most successful approach to designing short-channel MOSFETs is scaling. Starting with a long-channel device that delivers satisfactory performance, a miniaturized version is made by scaling the important features such that similar operation is maintained. In this way, short-channel effects can be minimized in miniaturized devices.

There are several approaches to scaling. One method is constant-field scaling in which the device dimensions and voltages are scaled by the same factor such that the electric field remains unchanged. However, some physical quantities do not scale in the same way. Generally, the two current components, drift and diffusion, do not scale identically. Though the diffusion current is relatively unimportant at long-channel lengths, it plays a strong role in short-channel devices. Therefore, for increasingly smaller devices, constant field scaling starts to fail. Other scaling methods have replaced constant field scaling, most notably empirical techniques. From these techniques, it can be demonstrated that miniaturized devices must have thinner gate oxides and either lower applied voltages or heavier doping concentrations in order to minimize short-channel effects.

The scaling process cannot be continued indefinitely, however, since it requires ever-decreasing gate oxide thicknesses. At some point, the gate oxide can no longer be thinned since it cannot be scaled to zero. Realistically, it is unlikely that the gate oxide can be thinned to less than 1 nm , at least if SiO_2 oxides are retained. Similarly, there are constraints on the maximum doping concentrations available. In addition, at smaller dimensions new quantum mechanical effects appear. Still other limits are imposed by the circuit and system requirements.

F. MOSFET Structures

A variety of modified basic MOSFET structures have been proposed to achieve high switching speed, improved current and voltage performance, and high packing density.

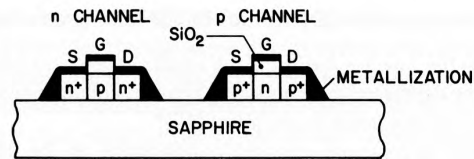


FIGURE 16 Silicon-on-sapphire with n - and p -channel MOSFETs.

1. Silicon-On-Sapphire Transistor

The silicon-on-sapphire transistor (SOS) is the most common of the various silicon on insulator (SOI) transistor structures that have been investigated. Figure 16 shows how the idea is implemented. A very thin silicon film is deposited by epitaxy on a highly polished sapphire crystal substrate. The film is etched to form p - and n -type small islands, which serve for the processing of the n - and p -channel MOSFETs, respectively. The sapphire single crystal is an excellent electric insulator and a good thermal conductor. Its primary purpose is to minimize extraneous parasitic capacitances that limit the conventional MOSFETs speed of response, and to provide nonjunction isolation between devices.

Ultra thin ($\sim 0.1\text{ }\mu\text{m}$) silicon films result in devices with low sensitivity to transient radiation effects. Perhaps the most important advantage of SOI devices is that short channel effects can be better controlled than in conventional devices.

2. Metal Nitride-Oxide Silicon Transistor

This device belongs to the class of metal insulator-oxide silicon (MIOS) transistors and is used mainly as a non-volatile element of read-only memories; it can store data even after power is removed.

The metal nitride-oxide silicon transistor (MNOS) Fig. 17 differs from the n -channel MOSFET in that the gate insulator is a composite of silicon dioxide and silicon nitride (Si_3N_4). The oxide layer is very thin ($2\text{--}5\text{ nm}$); the nitride layer, on the other hand, is comparatively thick ($30\text{--}50\text{ nm}$).

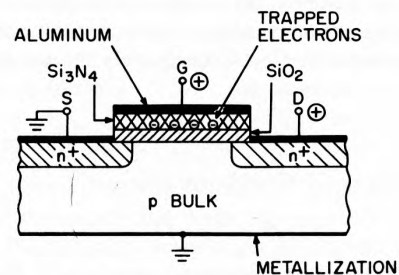


FIGURE 17 Schematic cross section of an MNOS transistor.

On the application of a positive gate voltage V_G the electrons from the bulk move by quantum mechanical tunneling through the thin oxide to the oxide–nitride interface, where they are trapped in localized states. The trapped negative charge adds up to the surface depletion region negative charge Q_{Dmax} , thereby increasing the threshold voltage magnitude. A negative gate voltage pushes the trapped electrons back into the silicon bulk thus lowering the threshold voltage. Since the MNOS transistor may be in one of these two states, it can serve as a binary element in a semiconductor memory. Switching from one state to the other may be effected millions of times before the device shows significant degradation. Metal nitride–oxide silicon transistors can be fabricated with either an n - or a p -type channel.

3. Floating-Gate Avalanche-Injection Transistor

The floating-gate avalanche-injection transistor (FAMOS) is another nonvolatile memory element. A floating (not electrically connected) polycrystalline gate is embedded in a SiO_2 layer, which is thick enough to prevent any quantum mechanical tunneling (Fig. 18). In order to charge the floating gate the drain voltage V_D is made high enough to cause the avalanche break-down of the drain junction. Some of the electrons generated and accelerated in the drain depletion region become energetic enough to reach the oxide, to penetrate it, and to drift in the field existing there to the floating gate, where they become trapped. The gate becomes negatively charged and causes the surface inversion of the n -type substrate, turning the device (a p -channel transistor is considered) on. The charge can be maintained for an indefinitely long time. Its erasure, which causes the device to revert to the off condition, is accomplished by irradiation by ultraviolet light or X-rays. Electrical erasure is made possible by the addition of another polysilicon gate deposited on the SiO_2 insulator above the floating gate. Such a structure is known as a stacked-gate avalanche-injection MOS (SAMOS).

4. Double-Diffused MOSFET

The double-diffused MOSFET (DMOS) is schematically represented in Fig. 19. A short and well-controlled channel

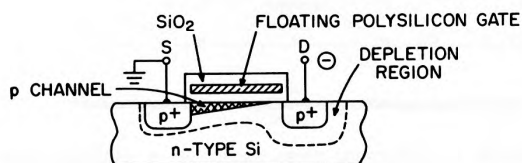


FIGURE 18 FAMOS transistor (schematic Representation).

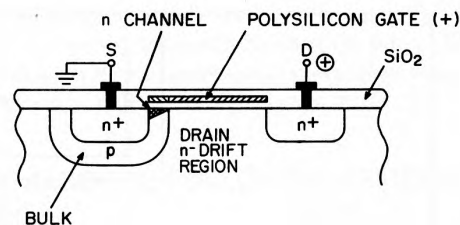


FIGURE 19 Schematic cross section of a horizontal DMOS.

length is obtained by the subsequent diffusions of the p -type bulk region into the n -type silicon crystal and the n -type source into the p -type bulk. Both diffusions take place through the same oxide opening. Similar structures can be obtained by double implantation (DIMOS).

The purpose of the relatively long n^- drift is to allow the drain depletion region to extend mostly into it. In this way the drain avalanche breakdown voltage can be greatly increased. Figure 19 shows the structure of a horizontal type of DMOS. For high-power devices, a similar structure is used, but often with a vertical n^- region.

5. Dual Gate MOSFETs

In addition to the usual gate formed on the top of the device, there can be another gate within the device. These devices are called dual gate MOSFETs. Both gates control the channel potential (Fig. 20). The two gates act to shield the source region of the channel from the action of the drain, providing a high degree of isolation of the source and drain. As a result, short-channel effects are reduced within the device.

6. SiGe MOSFETs

A strained Si layer can be formed using an Si/SiGe heterostructure. The electron mobility within strained Si is significantly higher than that of bulk silicon, and about a factor of five higher than that in Si inversion layers. The enhanced mobility within the strained silicon layer can be used to increase the carrier mobility within a MOSFET device. A higher mobility in turn leads to a higher transconductance and thus a higher frequency of operation of the device.

G. Power MOSFETs

Power MOSFETs are intended to switch or amplify currents of the order of tens of amperes and to support high circuit voltages. To do this the device should exhibit a very small on-resistance and a small voltage drop when the drain current is at its peak value. This is necessary

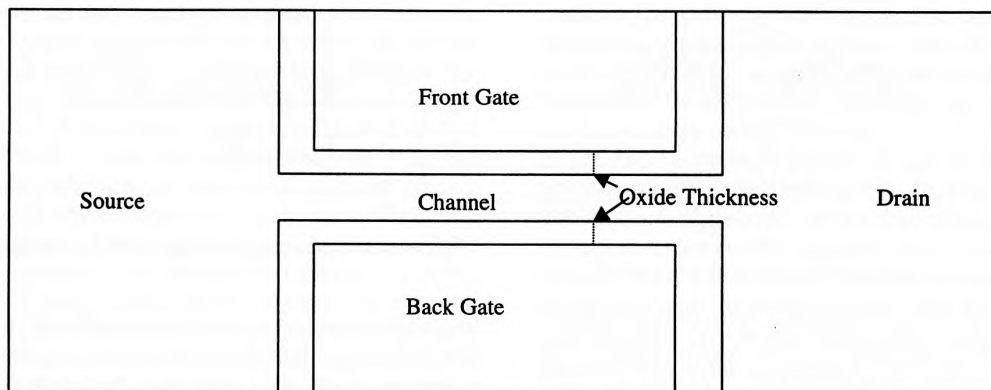


FIGURE 20 Sketch of the cross section of a dual gate MOSFET device. Notice that the channel is controlled by both the front and back gates.

to minimize power losses and keep the junction temperature below the maximum allowable limit of 150°C . The requirement of high drain breakdown voltage is in conflict with the low resistance requirement because it necessitates the use of high-resistivity silicon. For the same silicon crystal area, power MOSFETs are inferior to bipolar transistors from the point of view of on-resistance. On the other hand, they are superior as fast switches for such applications as switching-mode power supplies. As high input impedance devices they require less current drive, and their current gain does not degrade at high current levels. The MOSFET on-resistance has a positive temperature coefficient and therefore has a stabilizing effect, making the device resistant to "thermal runaway" instability. The so-called second breakdown, which severely limits the bipolar transistor safe operating area, is not as great a problem in the MOSFET.

Early power devices utilized a structure with a V-shaped gate obtained by anisotropic silicon etching in the manner described for the JFET in Section III.D. The device structure includes a vertical n^{-} drift region to support high drain voltages. Unfortunately, the V-shaped gate presented a breakdown problem so these devices have been almost entirely superseded by the DMOS (Section III.E.4) with a vertical drain.

The structure of a power MOSFET includes thousands of small MOSFETs integrated on the same silicon chip and operating in parallel to obtain a low on-resistance and therefore high current handling capability. Figure 21 shows the cross section of two n -channel enhancement-type devices forming two distinct cells, each consisting of a hexagonally shaped (as viewed from the top) n^{+} source and a similarly shaped polysilicon gate embedded in the oxide. An inversion layer is formed around the hexagonal source; the electrons from the n channel flow into the

n^{-} drift region, and then to the n^{+} substrate and drain contact. The vertical drain design permits excellent surface utilization and facilitates device heat sinking.

The source metallization shorts the source perimeter to the heavily doped p^{+} surface of the p^{-} region, providing a shunting resistance across the emitter (source) and the base (p bulk) of the parasitic transistor, reducing its current gain and the probability of second breakdown. The packing density of these structures can be as high as 10^5 – 10^6 cells per square centimeter with channels $1\text{ }\mu\text{m}$ long; total channel widths can be measured in tens of meters.

1. Conductivity-Modulated Field-Effect Transistor

The current capability of the power MOSFET is considerably extended in the conductivity modulated field-effect transistor (COMFET), which is also known as the insulated-gate transistor (IGT). In this device the n^{+} drain region of the power MOSFET (Fig. 21) is replaced by a very thin (2-nm) n^{+} film deposited on the p^{+}

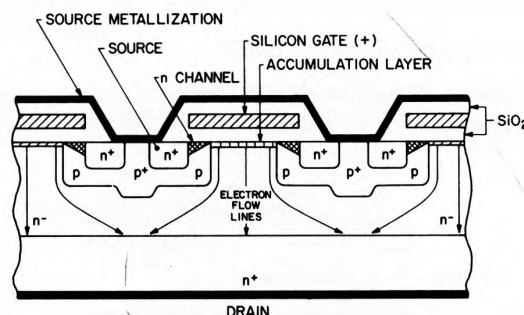


FIGURE 21 Schematic cross section of a multi-cell power MOSFET (vertical DMOS).

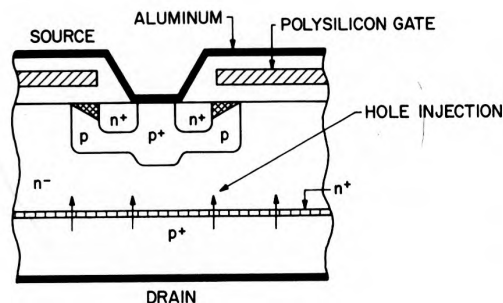


FIGURE 22 Schematic cross section of a COMFET.

substrate (Fig. 22). This creates in essence a four-layered ($n^+ - p - n^- - p^+$) thyristor-like structure activated by a MOSFET. When the device is turned on by the application of the positive gate voltage V_G greater than the threshold voltage V_T , electrons start flowing to the drain region, biasing the drain $n^- - p^+$ junction in the forward direction. The drain (the anode of the thyristor) consequently injects holes into the n^- drift region. To maintain charge neutrality, electrons enter the n^- region to balance the positive charge of holes, thereby increasing the n^- -region conductivity and decreasing the device on-resistance by a factor of 10 as compared with that of the MOSFET of the channel width Z .

The purpose of the p^+ region in each cell's center combined with the metallization shorting the n^+ and p^+ regions is to reduce the n^+ emitter into the p base (bulk) and the current gain α_1 of the $n^+ - p^+ - n^-$ transistor. The 2-nm-thick, heavily doped n^+ layer on top of the p^+ drain region lowers the emitter efficiency of the $p^+ -$

$n^+ - n^-$ junction, and therefore the current gain α_2 of the $p^+ - n - p^+$ transistor. Keeping the sum of the two current gains $\alpha_1 + \alpha_2$ below unity prevents latching of the $n - p - n - p$ device at high current levels.

SEE ALSO THE FOLLOWING ARTICLES

BIPOLAR TRANSISTORS • INTEGRATED CIRCUIT MANUFACTURE • OPTICAL AMPLIFIERS (SEMICONDUCTOR) • SEMICONDUCTOR ALLOYS • THIN FILM TRANSISTORS

BIBLIOGRAPHY

- Blicher, A. (1981). "Field-Effect and Bipolar Power Transistor Physics," Academic Press, New York.
- Blicher, A. (1982). "Physics of semiconductor power devices," *Rep. Prog. Phys.* **45**(4). The Institute of Physics, London, England.
- Brennan, K. F. (1999). "The Physics of Semiconductors with Applications to Optoelectronic Devices," Cambridge University Press, Cambridge.
- Brews, J. R. (1981). "Physics of the MOS transistors," In "Silicon Integrated Circuits," (D. Kahng, ed.), Academic Press, New York.
- Milnes, A. G. (1980). "Semiconductor Devices and Integrated Electronics," Van Nostrand Reinhold Company, New York.
- Nicollian, E. H., and Brews, J. R. (1982). "MOS (Metal-Oxide-Semiconductor) Physics and Technology" Wiley, New York.
- Richman, P. (1973). "MOS Field-Effect Transistors and Integrated Circuits," Wiley, New York.
- Sze, S. M. (1981). "Physics of Semiconductor Devices," 2nd ed., Wiley (Interscience), New York.
- Tsividis, Y. (1987). "Operation and Modeling of the MOS Transistor," McGraw-Hill, New York.
- Zambuto, M. (1989). "Semiconductor Devices," McGraw-Hill, New York.