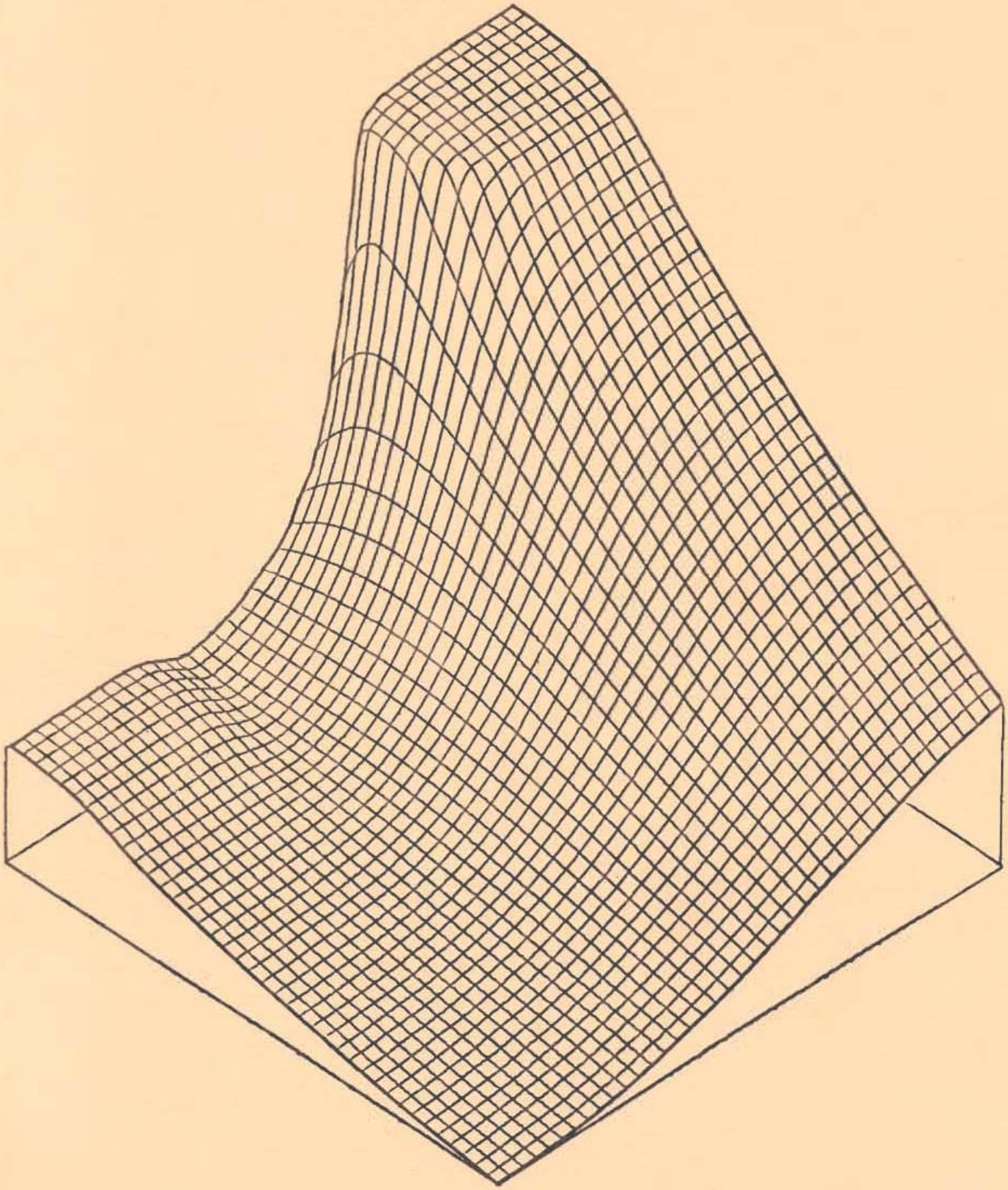
TWO DIMENSIONAL MODELING OF MOS TRANSISTORS

SIEGFRIED SELBERHERR



TECHNICAL UNIVERSITY OF VIENNA

VIENNA, AUSTRIA

ENGLISH TRANSLATION

OF

TWO DIMENSIONAL MODELING

OF MOS TRANSISTORS

BY

SIEGFRIED SELBERHERR

SOLD TO: _____

_____

_____

Additional copies may be ordered from:

SEMICONDUCTOR PHYSICS, INC.
639 Meadow Grove Place
Escondido, CA 92027
(619) 741-3360

PREFACE TO THE ENGLISH TRANSLATION

This English translation of Professor Selberherr's Ph. D. thesis was undertaken because I judge that the two dimensional MOS transistor simulation program MINIMOS which is described herein will become the "SPICE" of MOS device simulation in the eighties. MINIMOS shares many of the same virtues of SPICE when it was first released by U.C. Berkeley in the early seventies. The program came at a time when the need for such a tool had become critical. No other reliable, general purpose, two dimensional MOS transistor simulator is readily available. MINIMOS is free just for the asking and is already in use in a large number of university and industrial laboratories.

Because of the ready availability and widespread use of MINIMOS there will result a worldwide community of device researchers and technology developers who will rely completely upon MINIMOS for their device simulation capability ... much in the same way as a large fraction of the integrated circuit community has relied upon SPICE for circuit simulation capability during the seventies.

MINIMOS has demonstrated itself as being a reliable and accurate simulator of small dimension MOS transistors. However, the only detailed documentation with respect to the philosophy behind and the form of the internal structure of MINIMOS is contained in Professor Selberherr's Ph.D. thesis which is written in the German language.

Since a vast number of the users of MINIMOS are native English speaking, a great void exists which this English translation is meant to fill. This translation has been completed by Semiconductor Physics, Inc., a consulting firm which specializes in device physics, modeling and characterization. The translation is meant to be a one-to-one representation of Professor Selberherr's original work.

The translation was carried out not by professional language translators, but by native English speaking device physicists whose own backgrounds coincide with the topic of the thesis ... thus the accuracy of the translation has benefitted. Furthermore, Professor Selberherr has kindly consented to proofread the entire translation and provide the necessary corrections. The equations have been largely photocopied from the original text in order to preserve accuracy.

I am indebted to Professor Selberherr for having granted permission for this translation to be commercially distributed and for having provided high quality original size figures.

May 30, 1982

Escondido, California

Jim Smith, President

SEMICONDUCTOR PHYSICS, INC.

Abstract of the

D I S S E R T A T I O N

"Two-dimensional Modeling of MOS Transistors"

In fulfillment of the requirements
for the Degree of Doctor of Technical Science

Submitted to the Technical University of Vienna
by

Dipl.-Ing. Siegfried Selberherr
A-3441 Dietersdorf 61

Vienna, January 1981

Acknowledgment

I give special thanks to my teacher, Professor Dr. Hans Pötzl, for his great interest in the accomplishment of this work and for his helpful suggestions during countless discussions. Furthermore I thank all of my colleagues at the "Institute for Physical Electronics" for the friendly atmosphere, especially Dipl.-Ing. Alfred Schültz, without whose assistance the completion of the computer programs would have been greatly delayed.

I am deeply indebted to Siemens of Munich for the test transistors which were at my disposal.

My colleagues and friends at the "Institute for Numerical Mathematics", especially Professor Dr. Richard Weiß , Dipl.-Ing. Christian Ringhofer and associate Professor Dr. Christoph Überhuber, deserve my sincere thanks for numerous insightful conversations.

Finally I would especially like to thank the computer center of the Technical University of Vienna, Dipl.-Ing. Johannes Demel for his worthwhile suggestions with programming problems, Dipl.-Ing. Dieter Schornböck for generously providing computer time and Dipl.-Ing. Friedrich Blöser for the friendly help with the generation of graphics.

This work was supported by the "Funds for the Advancement of Scientific Research" (project no. S22/11).

Abstract

Because very large scale integration uses MOS transistors of very small dimensions it is necessary to use computer aided modeling in order to understand their behavior. There is no analytic model in the current literature which can adequately explain the short channel effects in MOS transistors.

This work presents a two-dimensional MOS transistor model which is based only upon fundamental physical principles, and a practically oriented computer program was developed which numerically solves the model equations. The optimization of the models by appropriate assumptions is illustrated and thereby the scope of its validity is defined. Physical parameters contained in the model, as for example the carrier mobility are described with the help of a set of mathematical equations, whose realistic details are discussed. The solution of the model equations, coupled non-linear partial differential equations, is obtained by the application of numerical mathematical procedures and therewith insight with regard to their fundamental functional form is gained.

The functional efficiency of the implemented computer program is demonstrated by selected examples. It is shown, that all typical effects in miniature MOS transistors can be satisfactorily explained with the help of a two-dimensional model. An inverter circuit with submicron transistors was analyzed and examined for correct operation. Further a sensitivity analysis of a modern transistor technology is described; a strategy for the determination of the limits of the miniaturization and an estimate of reproducibility in an established process is given. It is to be expected that the developed computer program will enable faster and improved development of modern MOS transistors in regard to very large scale integrated circuits.

TABLE OF CONTENTS

*) not included in the present draft of the dissertation

1. Introduction

In the 20 years since 1960 when Kahng and Atalla /68/demonstrated the first functional MOS transistor, a nearly inconceivable development of these devices has occurred. Today in the decade of very large scale integrated circuits, electrical engineering without the MOS transistor is totally inconceivable. The very large scale integration (VLSI), which is mentioned here, is a technology which also absolutely requires computer aided simulation of its devices. The design of modern MOS transistors by purely experimental methods, experience and analytic models would be extremely time consuming, often very expensive and sometimes impractical from a technical viewpoint. The thus far published analytical models of MOS transistors depend upon certain assumptions, which impose certain physical restrictions, such that, in general only a limited ability to analyze and predict transistor characteristics can be achieved. Especially because of the advancing rate of miniaturization, these simple models are losing their usefulness. In order to describe the modern MOS transistor in a useful way, one is compelled to use numerical models with increased accuracy and without fundamental physical restrictions.

About 15 years ago there appeared the first, yet one dimensional, model which did not use the regional approximation as its foundation, but instead used the fundamental semiconductor transport equations /134/. The first consistent simulation using such a model was published by Gummel /58/ for the bipolar transistor. There followed a flood of analogous work with refinements and improvements in one way or another. De Mari e.g. simulated the static /35/ and dynamic /34/ operation of the P-N diode and through this work worthwhile mathematical and physical suggestions were gained.

At the end of the 1960's these one-dimensional models were extended to two coordinates. Slotboom investigated the bipolar transistor in one and two dimensions /118/, Kennedy began with the JFET /72/, Dubock performed a two dimensional analysis of the diode /41/ and transistor /40/ and Loeb attempted a two dimensional calculation for the MOS transistor. Today there exists a wide spectrum of literature on the modeling of different devices with their widely varied structures which is obvious from the bibliography which does not

claim to be complete.

The development of a two-dimensional model necessarily requires implementation in a computer program. Up until now the programs which have been developed have not been widely available. The authors had to deal with numerical stability, limited flexibility, large memory requirements and computation time requirements or low throughput.

Likewise, in this work a computer program was developed, called MINIMOS; a program aimed at consistent, numerical simulation of MINIature MOS transistors. The greatest value of the program does not lie solely in its physical foundation but also in its flexibility, modularity, dynamic memory management and portability. It should become a tool not only of academic interest but it can also be used for modeling modern transistors. The judgement, of whether or not this has been well done, naturally cannot be predicted here; an evaluation can be found in /51/ and many international institutions have indicated their interests and MINIMOS can be obtained in academic exchange for practical usage. MINIMOS should represent, with the feedback from the international users, a massive cornerstone for further modeling and design work, in that the basic understanding of the behavior of MOS transistors will be enhanced.

Chapter 2 of this work deals with the physical foundations of a two-dimensional MOS transistor model. The fundamental semiconductor equations and the permissible simplifications will be discussed and the modeling of physical parameters explained.

The theme of chapter 3 is the numerical point of view of the MOS transistor. The transformation of the fundamental equations to a form suitable for numerical solution will be discussed and the method of solution explained.

In chapter 4 it will be attempted, on the one hand, to demonstrate by examples, the functional capability of the computer program, and on the other hand, to confirm the correctness and plausibility of the underlying physical models. The selected examples should clarify the two-dimensional model, especially its breadth of applicability.

2. The Physical Model

In this chapter the necessary physical assumptions for a two-dimensional numerical MOS transistor model will be discussed.

Section 2.1 deals with the correct formulation of the fundamental semiconductor equations. An attempt will be made to define the scope of validity of these fundamental equations. The simplification of the general equations will be illustrated with appropriate assumptions and justified with physical arguments. The model equations which are valid for the MOS transistor (simplified fundamental semiconductor equations) will be transformed to dimensionless form and summarized.

Section 2.2 deals with the modeling of the physical parameters of the simplified fundamental equations. Special attention will be given to the simulation of the impurity distribution and the mobility. A previously unpublished model for surface scattering will be presented and the physical reasoning upon which the model is based will be explained.

2.1 The Fundamental Equations

In order to accurately analyze an arbitrary semiconductor structure under all kinds of operating conditions the classical fundamental semiconductor equations must be solved. These are contained in the following five partial differential equations.

$$\operatorname{div}\, \varepsilon\, \operatorname{grad}\, \psi = -q\cdot(p-n+N_D^+-N_A^-) \qquad (2.1\text{-}1)$$

$$\operatorname{div}\, \vec{J}_n - q\cdot(\partial n/\partial t) = q\cdot R \qquad (2.1\text{-}2)$$

$$\operatorname{div}\, \vec{J}_p + q\cdot(\partial p/\partial t) = -q\cdot R \qquad (2.1\text{-}3)$$

$$\vec{J}_n = -q\cdot(\mu_n\cdot n\cdot \operatorname{grad}\, \psi - D_n\cdot \operatorname{grad}\, n) \qquad (2.1\text{-}4)$$

$$\vec{J}_p = -q\cdot(\mu_p\cdot p\cdot \operatorname{grad}\, \psi + D_p\cdot \operatorname{grad}\, p) \qquad (2.1\text{-}5)$$

Equation (2.1-1) is Poisson's equation, which characterizes the charge distribution in the semiconductor. Equation (2.1-2) describes the balance of the source and sink of electron current and eq. (2.1-3) gives the analogous relationship for hole current. These are called the continuity equations. The magnitude and direction for electron current are given by (2.1-4) and for holes by (2.1-5). This set of equations which describe the transport phenomenon in a semiconductor device was first given in closed form by Van Roosbroeck /134/ in 1950. It is to be emphasized, however, that these equations do not describe degenerate effects. In /86/, /92/, and /136/ the modifications necessary in order to take into account degeneracy (the breakdown of Boltzmann statistics and the variation of the band edge as well as the variation of the band gap) for eqs. (2.1-4) and (2.1-5) are discussed. These modifications are, however, in part not simple and lead to special problems in the boundary conditions /94/ and /135/. A consistent, but only one-dimensional, model of the degeneracy phenomenon was first published in 1979 /93/. In the relatively low doped (less than 10^{17} cm^{-3}) channels of MOS tranistors, where the current transport occurs, the degeneracy phenomenon plays absolutely no role and will not be considered here.

2.1.1 The Chosen Assumptions

Some assumptions in the presented model have been touched upon which significantly simplify the solution of the equations without a considerable

loss of accuracy. The assumptions should in no way alone remove the difficulties as one might falsely conclude from the last sentence. On the contrary, in the first place, the throughput and computational speed for the solution of the fundamental equations by an obviously necessary computer program should be raised if the program is to become a worthwhile tool. The program developed in the scope of this work shall perform in developmental efforts, sometimes with very many variables, not only for the infrequent simulation of academic interest, but also for the experienced transistor designer.

* Only a static solution is sought. This assumption is of fundamental importance, thereby the order of the partial differential equations is reduced. When considered from a mathematical point of view, this leads to a significant change in the formulation of and insight into the problem. By way of the suppression of the time dependent term a parabolic problem is converted to an elliptic problem, whereby another point of view must be adopted. Contributions toward the transient solution of the semiconductor equations have indeed already been published e.g. /10/, /30/, /84/, /100/, and /102/, however the authors themselves acknowledge that these solutions are only of academic interest and the programs developed by these authors have exhibited limited flexibility and therefore are of little practical value. However, relatively new theoretical work /76/, /91/ has given hope that in the near future a useful model for a two dimensional transient analysis will be developed.

$$\partial n/\partial t = 0 \qquad (2.1\text{-}6)$$

$$\partial p/\partial t = 0 \qquad (2.1\text{-}7)$$

* On the grounds of the lattice structure of silicon and the amorphous nature of oxide, their dielectric constants are isotropic. The dielectric constants in equation (2.1-1) can be taken outside the divergence operator.

$$\varepsilon_{si} = \text{const} \qquad (2.1\text{-}8)$$

$$\varepsilon_{ox} = \text{const} \qquad (2.1\text{-}9)$$

* Total ionization of the impurities will be assumed, which is justified for the temperature range of 250K to 450K to which this model is limited /121/.

$$C = N_D - N_A = N_D^+ - N_A^- \qquad (2.1\text{-}10)$$

* Degeneracy phenomena, as has already been mentioned, will not be considered. The intrinsic carrier concentration will be taken as constant.

$$n_i = \text{const} \qquad (2.1\text{-}11)$$

* Majority carrier current is neglected. This assumption represents the most significant limitation. The a priori assumption of negligible majority carrier current prevents a direct and consistent calculation of substrate current and an analysis of the breakdown behavior. However, though an integration of the ionization rate throughout the entire transistor, a satisfactory evaluation of the substrate current can be made /131/, and the onset of avalanche breakdown under high field conditions can be found. In principle this assumption is justified, inasmuch as two of the five fundamental equations become trivial and do not need to be solved.

$$\vec{J}_p = 0 \quad \text{(for n-channel transistors)} \qquad (2.1\text{-}12)$$

$$\vec{J}_n = 0 \quad \text{(for p-channel transistors)} \qquad (2.1\text{-}13)$$

* The temperature throughout the entire transistor is constant and can be varied over the interval of 250K to 450K.

$$T = \text{const} \qquad (2.1\text{-}14)$$

* The carrier distribution is described by Boltzmann statistics. This assumption states that heavy doping and degeneracy effects need not be considered and therefore are not a problem.

$$n = n_i \cdot e^{(\psi - \phi_n)/U_T} \qquad (2.1\text{-}15)$$

$$p = n_i \cdot e^{(\phi_p - \psi)/U_T} \qquad (2.1\text{-}16)$$

* The validity of the Einstein-Nernst relation is assumed. The importance of this assumption is tied to Boltzmann statistics and the original form of the fundamental semiconductor equations, in that these formulations do not distinguish between the electron temperature and the lattice temperature.

$$D_n = \mu_n \cdot U_T \qquad (2.1\text{-}17)$$

$$D_p = \mu_p \cdot U_T \qquad (2.1\text{-}18)$$

* All contacts are considered to be ohmic. The space charge vanishes at the contact, and the carrier distribution is in thermal equilibrium.

2.1.2 The Model Equations

In consideration of the chosen assumptions found in the last section the fundamental equations are simplified considerably. But before the simplified equations are summarized, a normalization into dimensionless form will be carried out following De Mari /35/. The normalization constants for electric charge, the dielectric constant, the diffusion constant, the flux and the potential will be determined a priori, whereupon everything else in the equations is constrained to follow. The important normalization factors are listed in figure 2.1-1.

Dimension	Units	Normalization factor
Electric charge	As	q
Diffusion constant	cm^2/s	$D_0 = 1\ cm^2/s$
Dielectric constant	As/Vcm	ε_{si}
Flux	cm^{-3}	n_i
Potential	V	U_T
Length	cm	$x_0 = ((\varepsilon_{si} * U_T)/(q * n_i))^{\frac{1}{2}}$
Time	s	x_0^2/D_0
Mobility	cm^2/Vs	D_0/U_T
Current density	A/cm^2	$J_0 = q * D_0 * n_i / x_0$

Figure 2.1-1: The normalization factors

The postulates (2.1-12) and (2.1-13) necessarily require a distinction between the equations for n-channel and p-channel transistors. Consequently for n-channel transistors the following equations hold:

$$\text{div grad } \psi = e^{\psi-\phi_n} - e^{\phi_p-\psi} - C \qquad (2.1\text{-}19)$$

$$\text{div } \vec{J}_n = R$$

$$\vec{J}_n = -\mu_n \cdot n \cdot \text{grad } \phi_n$$

$$\phi_p = \text{const} \quad (\text{which implies: } \vec{J}_p = 0)$$

and for the p-channel transistor the following equations hold:

$$\text{div grad } \psi = e^{\psi-\phi_n} - e^{\phi_p-\psi} - C \qquad (2.1\text{-}20)$$

$$\text{div } \vec{J}_p = -R$$

$$\vec{J}_p = -\mu_p \cdot p \cdot \text{grad } \phi_p$$

$$\phi_n = \text{const} \quad (\text{which implies: } \vec{J}_n = 0)$$

Equations (2.1-19) and (2.1-20) actually describe a coupled system of two nonlinear partial differential equations, which can only be solved numerically. It should probably also be mentioned here, that all quantities in both of these systems of equations appear dimensionless, which implies normalized variables. For the purpose of clarity these were not characterized by explicit indices.

2.2 The Physical Parameters

The model equations derived in the last section (2.1-19) and (2.1-20) contain several physical parameters whose modeling will be discussed next. The importance of these parameters may not be underrated, in that their accuracy directly determines the quantitative validity of the total simulation results.

* The thermal voltage U_T is the simplest parameter. It is only required for normalization (see section 2.1.2) and is only dependent upon the simulation temperature.

$$U_T = k \cdot T/q \quad (V) \qquad (2.2\text{-}1)$$

* The intrinsic carrier concentration is modeled as being only temperature dependent. The formula used is very simple and does not account for the temperature dependence of the band gap and only roughly accounts for the temperature dependence of the band edge /50/. These temperature dependencies however are not necessary because the intrinsic carrier concentration is only required for normalization and furthermore the temperature is a global quantity. The narrowing of the band gap due to heavy doping /89/, /119/ can also be neglected.

$$n_i = 3.88 \cdot 10^{16} \cdot T^{1.5} \cdot e^{-7000/T} \quad (cm^{-3}) \qquad (2.2\text{-}2)$$

* The thermal generation will be simulated by means of simple Shockley-Read-Hall levels. The neglect of the majority carrier current in general appears to give these levels no meaning. For the intrinsic current transport in the channels of MOS transistors it is also true that on the grounds that the majority carrier current is absent, no recombination is possible. Without thermal generation, which would be naturally described by these levels, an absolutely unrealistic carrier depletion arises in the reverse biased drain/substrate diode. These levels will be chosen for stabilization in order to eliminate a numerical problem which may arise due to an unrealistically low density. The recombination centers are assumed to be in the middle of the forbidden band. The lifetimes will not be modeled /31/, but will instead be

assumed to be constant.

$$R = \frac{p \cdot n - n_i^2}{\tau_p \cdot (n + n_i) + \tau_n \cdot (p + n_i)} \quad (cm^{-3}s^{-1}) \qquad (2.2\text{-}3)$$

The remaining parameters, the doping profile and the mobility, are of primary importance in the behavior of modern MOS transistors and must be modeled with greater effort.

2.2.1 The Doping Profile

The most important input parameter for the simulation of miniature transistors is the doping profile. International research in the area of two-dimensional doping profiles is just now beginning. Because of the difficulties hidden in the modeling refinements of the diffusion of the impurities; the modeling of the diffusion constant; the effects due to interaction of different dopants, the oxide growth and similar questions, an interfaculty cooperation is needed in order to achieve concrete progress in these areas. Interesting results and aspects can be inferred from /106/ about two-dimensional implantation effects and from /107/, /130/, and /141/ about two dimensional diffusion effects. The possibility of obtaining closed form analytic solutions by realistic model refinements is as small for the diffusion problem as it is for the electrical transport problem. The importance of coupling between two-dimensional process simulation and two-dimensional transistor modeling will surely arise in the highest degree in the near future because of progressive miniaturization /20/, and will provide great stimulus for scientific contributions in many places.

In MINIMOS there are two principle possibilities available for use in specifying the doping profile. SUPREM, the Stanford University Process Engineering Model Program /6/ can be used to calculate, with very good accuracy, a one-dimensional channel profile and source/drain profile, whereby in the lateral direction an adjustment with equation (2.2-4) will be made.

$$C(x,y) = C((y^2+\max(x/f,0)^2)^{1/2}) \qquad (2.2\text{-}4)$$

In the above equation $y = 0$ represents the interface; $x = 0$ denotes the edge of the gate mask at the source; the oxide mask lies in the negative x

direction. Equation (2.2-4) permits control of the underdiffusion with the help of the parameter 'f', which in practical cases lies between 0.5 and 0.9. This equation does not take into account the outdiffusion in the positive x direction.

The second facility for the specification of the doping profiles uses an analytic approach, which in many cases provides sufficient accuracy. This approach is based principally upon the work of Lee /78/, /79/.

A predeposition can be simulated with equation (2.2-5). The parameters required here are the surface concentration N_s, the diffusion time t, and also the diffusion constant D for the doping element and the diffusion temperature. A predeposition can only be simulated for the source/drain regions, because it is only meaningful in those regions. Similar approaches for modeling the predeposition are given in /19/ and /70/.

$$ld = 2 \cdot \sqrt{D \cdot t} \qquad (2.2\text{-}5)$$

$$C_p(x,y) = 0.5 \cdot N_s \cdot e^{-(y/ld)^2} \cdot \mathrm{erfc}(x/ld)$$

Ion implantation and diffusion can be simulated with the set of equations (2.2-6). The parameters which must be specified are the dopant element (which affects R_p, ΔR_p and D), the implantation dose "DOSE", the implantation energy (which defines R_p and ΔR_p), the oxide thickness T_{iox} (which influences R_p), the diffusion time t, and the diffusion temperature (hidden in the diffusion constant). When double channel implantations are performed, the individual profiles are superimposed.

$$ld = 2 \cdot \sqrt{D \cdot t} \qquad (2.2\text{-}6)$$

$$a = (2 + (ld/\Delta Rp)^2)^{-1/2}$$

$$K(y) = e^{-(a \cdot (Rp-y)/\Delta Rp)^2} \cdot \mathrm{erfc}(-a \cdot ((Rp/\Delta Rp) + \sqrt{2} \cdot y/ld))$$

$$C_i(x,y) = (a/(4 \cdot \Delta Rp \cdot \sqrt{\pi})) \cdot Dose \cdot (K(y)+K(-y)) \cdot \mathrm{erfc}(x/ld)$$

It is further taken that the source/drain profile and the channel profile are simply superimposed, that is, interaction of the diffusions is neglected.

The diffusion constants are taken as constant in (2.2-5) and (2.2-6). This is justifiable for the diffusion of the channel profile (relatively low surface doping), but not for the source/drain diffusions. Because, at the present, no closed form analytic representation for diffusion with a nonconstant diffusion coefficient is known. No other alternative is available. One can, however, with some degree of skill with an arbitrary time or diffusion temperature, cause the pn transition region to appear as if the diffusion were done with dopant dependent and field dependent diffusion coefficients. The diffusion coefficients will be calculated by way of the traditional exponential equation and the data from figure 2.2-1.

$$D = D_0 \cdot e^{T_a/T} \qquad (2.2\text{-}7)$$

Element	$D_0/(cm^2s^{-1})$	$T_a/(K)$
B	0.5554	$-3.975 \cdot 10^4$
P	3.85	$-4.247 \cdot 10^4$
SB	12.9	$-4.619 \cdot 10^4$
A	24.	$-4.735 \cdot 10^4$

Figure 2.2-1: The diffusion constants.

Furthermore, the parameters R_p and ΔR_p, the projected range and the associated standard deviation, which appear in equations (2.2-5) and (2.2-6) must also be modeled. These parameters are tabulated in /55/ as functions of the implantation energy according to the LSS theory. That table, even though relatively coarse, can only be implemented in a computer program in an unwieldly fashion, and is interpolated by a best-approximation polynomial for the projected range and the standard deviation, thereby minimum memory storage requirements and very fast output can be achieved /115/. R_p and ΔR_p, take on the following forms.

$$Rp = \sum_i^n a_i \cdot x^i \qquad (2.2\text{-}8)$$

$$\Delta Rp = \sum_i^n b_i \cdot x^i \qquad (2.2\text{-}9)$$

In (2.2-8) x is understood to be the implantation energy. The rank n of these polynomials lies in each case, according to the implantation element, between 2 and 5. The polynomial coefficients for (R_{psi}) and (ΔR_{psi}) the projected range and standard deviation in silicon are given in figures 2.2-2 and 2.2-3 respectively and in figure 2.2-4 for (R_{pox}) the projected range in silicon dioxide.

Element	B	P	SB	A
a_1	$3.338 \cdot 10^{-3}$	$1.259 \cdot 10^{-3}$	$8.887 \cdot 10^{-4}$	$9.818 \cdot 10^{-4}$
a_2	$-3.308 \cdot 10^{-6}$	$-2.743 \cdot 10^{-7}$	$-1.013 \cdot 10^{-5}$	$-1.022 \cdot 10^{-5}$
a_3		$1.290 \cdot 10^{-9}$	$8.372 \cdot 10^{-8}$	$9.067 \cdot 10^{-8}$
a_4			$-3.056 \cdot 10^{-10}$	$-3.442 \cdot 10^{-10}$
a_5			$4.028 \cdot 10^{-13}$	$4.608 \cdot 10^{-13}$

Figure 2.2-2: Coefficients for R_p in silicon.

The dimension of the coefficients a_i, b_i in figures 2.2-2 through 2.2-4 is ($\mu m/keV^i$).

Element	B	P	SB	A
b_1	$1.781 \cdot 10^{-3}$	$6.542 \cdot 10^{-4}$	$2.674 \cdot 10^{-4}$	$3.652 \cdot 10^{-4}$
b_2	$-2.086 \cdot 10^{-5}$	$-3.161 \cdot 10^{-6}$	$-2.885 \cdot 10^{-6}$	$-3.820 \cdot 10^{-6}$
b_3	$1.403 \cdot 10^{-7}$	$1.371 \cdot 10^{-8}$	$2.311 \cdot 10^{-8}$	$3.235 \cdot 10^{-8}$
b_4	$-4.545 \cdot 10^{-10}$	$-2.252 \cdot 10^{-11}$	$-8.310 \cdot 11^{-10}$	$-1.202 \cdot 10^{-10}$
b_5	$5.525 \cdot 10^{-13}$		$1.084 \cdot 10^{-13}$	$1.601 \cdot 10^{-13}$

Figure 2.2-3: Coefficients for ΔR_p in silicon.

Element	B	P	SB	A
a_1	$3.258 \cdot 10^{-3}$	$9.842 \cdot 10^{-4}$	$7.200 \cdot 10^{-4}$	$7.806 \cdot 10^{-4}$
a_2	$-2.113 \cdot 10^{-6}$	$-2.240 \cdot 10^{-7}$	$-8.054 \cdot 10^{-6}$	$-7.899 \cdot 10^{-6}$
a_3			$6.641 \cdot 10^{-8}$	$7.029 \cdot 10^{-8}$
a_4			$-2.422 \cdot 10^{-10}$	$-2.653 \cdot 10^{-10}$
a_5			$3.191 \cdot 10^{-13}$	$3.573 \cdot 10^{-13}$

Figure 2.2-4: Coefficients for R_p in silicon dioxide.

When implanting through a protective oxide, the actual range must be reduced with equation (2.2-10), in which the projected ranges for silicon (R_{psi}) and for oxide (R_{pox}) and the oxide thickness (T_{iox}) are the required parameters /108/.

$$Rp = Rp_{si} \cdot (1 - T_{iox}/Rp_{ox}) \qquad (2.2\text{-}10)$$

2.2.2 The Mobility

The mobility is the most complex parameter in the fundamental semiconductor equations. Its modeling is of eminent importance, since any error in the mobility immediately influences the current density distribution in the transistor. A comparison of simulation results and measurements is as yet only possible by way of current characteristics, whereby the simulation is calculated pointwise through the integration of the current density, so that errors in the current density naturally cause a direct error in the current.

For a correct modeling of the mobility, the different underlying physical mechanisms must be taken into consideration. The basic mobility in high purity, field free silicon is determined by lattice scattering. In the following, this basic mobility will always be designated with μ_L (lattice). It is only temperature dependent and it can be modeled in a relatively simple way by way of a power law /64/, /80/. Equation (2.2-11) gives the applicable model of the basic mobility, whereby the indices n and p in the coefficients signify the identification of electrons and holes respectively.

$$\mu_L(T) = A \cdot T^{-\gamma} \quad (cm^2/Vs) \qquad (2.2\text{-}11)$$

$A_n = 7.12 \cdot 10^8$ $\qquad$ $A_p = 1.35 \cdot 10^8$

$\gamma_n = 2.3$ $\qquad$ $\gamma_p = 2.2$

Only the basic mobility has been described so far, with mention already made of the scattering in high purity, field free silicon. Through the existence of impurities this basic mobility will be decreased by way of two dimensional scattering of the free charge carriers due to impurities. This process is temperature dependent as is lattice scattering. A whole range of formulations have been published in order to model these effects. Many are based upon theoretical considerations /36/, /80/; quite a lot however are heuristic in nature but they are very good when their quantitative accuracy is checked experimentally /5/, /18/, /110/. The heuristic formulations have in general a substantially more simple structure with equal accuracy, so that in this work they will be given preference. Equation (2.2-12) will be used for concrete modeling of the temperature dependence of the composite lattice and impurity scattering mobility μ_{LI} in doped, but field free silicon.

$$\mu_{LI}(N,T) = \mu_L(T) \cdot a + \mu_{min} \cdot (1 - a) \quad (cm^2/Vs) \qquad (2.2\text{-}12)$$

$$a = \frac{1}{1 + (T/300)^{\delta} \cdot (N/N_0)^{\alpha}}$$

$N = N_D^+ + N_A^-$

$\mu_{minn} = 55.24$ $\qquad$ $\mu_{minp} = 49.7$

$\delta_n = -3.8$ $\qquad$ $\delta_p = -3.7$

$\alpha_n = 0.73$ $\qquad$ $\alpha_p = 0.7$

$N_{0n} = 1.072 \cdot 10^{17}$ $\qquad$ $N_{0p} = 1.606 \cdot 10^{17}$

Because in the MOS transistor the current flows mainly at the surface of the silicon, in the boundary layer which lies inside the inversion channel, a further scattering mechanism, the surface scattering, absolutely must be taken into account. It is established, however with great regret, that a model with a physical foundation is impossible, because theoretically this scattering process is not sufficiently understood. From this point on a further description of the surface scattering can only use heuristic arguments, which are based upon former intuitive reasoning. Regretably it is almost impossible to make a measurement of the surface mobility, which could directly verify a heuristic formula, because effectively one only measures the average mobility in the channel, however, a two dimensional mobility distribution is required for a two dimensional simulation. It is also safe to say, that the measurement of the effective mobility in the inversion channel exhibits great difficulty, which can only be mastered with satisfactory accuracy in outstanding laboratories /109/.

There is very little written about the heuristic considerations for models of the surface mobility /142/; the previously established considerations are simple and unsatisfactory. Experimental results are published in /124/, and future progress seems hopeful with respect to the modeling of the surface mobility.

From the above mentioned arguments a likewise heuristic formulation for a model of surface mobility and its correction with lattice and impurity mobility has been developed, with which plausible simulation results and a satisfactory accuracy can be obtained. With the help of the complete formula in (2.2-13) it is possible to include the gate controlled electric field dependence of the surface roughness scattering in the model. All effects describing the mobility which have been discussed up to this point are represented by the symbol μ_{LIS} (lattice, impurity, surface).

$$\mu_{LIS}(y,E_p,E_t,N,T) = \mu_{LI}(N,T)\cdot\frac{y+y_r}{y+b\cdot y_r} \quad (cm^2/Vs) \quad (2.2\text{-}13)$$

$$y_r = y_0/(1+E_p/E_{p0})$$

$$b = 2+E_t/E_{t0}$$

$$E_p = \max(0,(E_x\cdot J_x+E_y\cdot J_y)/(J_x^2+J_y^2)^{1/2})$$

$$E_t = \max(0,(E_x\cdot J_y-E_y\cdot J_x)\cdot J_x/(J_x^2+J_y^2))$$

$$y_{0n} = 5\cdot10^{-7} \qquad y_{0p} = 4\cdot10^{-7}$$

$$E_{p0n} = 10^4 \qquad E_{p0p} = 8\cdot10^3$$

$$E_{t0n} = 1.8\cdot10^5 \qquad E_{t0p} = 3.8\cdot10^5$$

The lattice/impurity mobility at the surface (y = 0) is reduced by the factor 1/b; and at a distance y_r it is reduced by the factor 2/ (1+b); and at greater distances from the surface it naturally follows that there is no reduction in the mobility. y_r represents a characteristic length, which describes the range of influence of the surface. This range is a function of E_p, the field strength component, which lies parallel to the current direction. The formulation for y_r produces a reduction in the range of the surface scattering by greater field strength parallel to the current direction, thereby velocity saturation appears, which will be discussed in the following. The remaining physical consideration is, that the charge carrier, which is moving at saturated velocity, experiences less influence due to the surface. The parameter b in (2.2-13) descibes the extent of influence of the surface scattering. It is a function of E_t, the projection of the field strength component normal to the current direction in the direction normal to the surface. The formulation for b rests upon the consideration, that the charge carriers are pressed against the surface by an electric field, which results in a greater scattering, such that a greater mobility reduction results.

The last relevant physical effect for modeling the mobility is the velocity saturation. For this effect there are no useful physical fundamental arguments of high accuracy, so that analogously to the treatment for the surface scattering a simple heuristic model must be derived. Formulations for

such models exist in the relevant literature in sufficient quantity /17/, /18/, /64/, /65/, /110/. The formulation used in this work is quoted from established literature in the form of a plausible common denominator. With the addition of the velocity saturation one obtains the formula (2.2-14), in which E_p, as in (2.2-13), is the field strength component in the current direction, and v_s is the saturation velocity. Formula (2.2-14) is considered to be a structure of a type of Mathiessen's rule /111/ with a correlation weight $(-\beta)$.

$$\mu_{tot}(y,E_p,E_t,N,T) = (\mu_{LIS}(\ldots)^{\beta}+(v_s/E_p)^{\beta})^{1/\beta} \qquad (2.2\text{-}14)$$

$$E_p = \max(0,(E_x \cdot J_x + E_y \cdot J_y)/(J_x^2+J_y^2)^{1/2})$$

$$v_{sn} = 1.53 \cdot 10^9 \cdot T^{-0.87} \qquad v_{sp} = 1.62 \cdot 10^8 \cdot T^{-0.52}$$

$$\beta_n = -2 \qquad \beta_p = -1$$

3. The Numerical Model

A central theme of the work presented here is the transformation of the physical model of the MOS transistor, which was established in the last chapter, into a numerical model. This numerical model was realized in the form of a computer program, which permitted the verification of the numerical model, as well as the physical model in wider scope. MINIMOS - so the program was named contains all of the necessary software for the solution of the semiconductor equations and the modeling of the physical parameters. In this chapter the numerical assumptions and background, that are necessary for the establishment of a program such as MINIMOS will be treated and discussed.

In section 3.1 the alternatives for the linearization of the fundamental equations will be given and the procedures implemented in MINIMOS will be explained.

Section 3.2 deals with the discretization of the fundamental equations. The discretization of a problem, which is presented as a transformation from an analytic formulation to a numerical formulation, will first be presented as carried out in general for the quasi-harmonic differential equation, a generalized type for the linearization of Poisson's equation and the continuity equation and it will then be specialized for these equations.

The solution of the discretized fundamental equations will be discussed in section 3.3 An overview of the available existing methods will be presented and their consequences singled out for the reader.

3.1 The Linearization of the Fundamental Equations

In the last chapter a system of nonlinear partial differential equations was presented, which describe the current transport in an MOS transistor, and which must be linearized before it can be solved numerically. The system of equations examined in the following are exclusively for N-channel transistors. The equations for the P-channel transistor are structurally identical and thereby all established considerations will be valid. For purposes of clarity this system is presented again here in normalized form:

$$\text{div grad } \psi = n - p - C \tag{3.1-1}$$

$$\text{div}(\mu_n \cdot n \cdot \text{grad}\,\phi_n) = (1-n\cdot p)/(\tau_p\cdot(n+1)+\tau_n\cdot(p+1)) \tag{3.1-2}$$

mit: $n = e^{\psi-\phi_n}$

$p = e^{\phi_p-\psi}$

$\phi_p = \text{const}$

$C = C(x,y)$

$\mu_n = \mu_n(x,y)$

$\tau_p = \text{const}$

$\tau_n = \text{const}$

The classical mathematical way to the solution of this kind of system is the use of a Newton procedure with eventual damping and/or extrapolation /99/, /104/ for the simultaneous solution of the complete system. A shortcoming of this method however is the considerable storage requirement which is required for the Jacobian matrix of the system. A more important advantage to be noted is that this procedure exhibits quadratic limiting convergence.

An alternative means for solving this system of equations is the application of a block nonlinear iteration procedure, by which a Newton-like formulation is not applied to the complete system. Instead a Newton-like formulation is established for each of the differential equations keeping constant the secondary independent variables for which an individual choice

must be found for each specific problem. From well known authors /15/, /30/ comparisons of both of these procedures have been made, and the conclusions conceded, that the block-nonlinear iteration procedure is preferred for a wide spectrum of applications. The procedures were first published in explicit form for application to semiconductor equations by Gummel /58/, who therein gave an intuitive and physically based derivation. In the relevant literature it is often designated as the Gummel algorithm.

A mathematical convergence proof for the application of the block-nonlinear iteration procedure to the semiconductor equations is not known and in considering the complexity of the problem, such a proof may not be trivial.

As a complement to the somewhat heuristic theory of the Gummel algorithm, the works /11/ and /127/ are mentioned, which however present only insignificant improvements.

The practical procedure followed is: One solves Poisson's equation with fixed φ_n and thereafter the continuity equation with constant ψ . This process is iteratively repeated until one has obtained a consistent solution for ψ and φ_n. The necessary accuracy is an important point to be stated for the solution of the individual equations. The accuracy is verified by beginning the iteration process with small accuracy requirements and successively increasing them during the iteration under consideration of the global errors from the right hand sides of (3.1-1) and (3.1-2).

One first applies the Newton-like formulation for the linearization of Poisson's equation (3.3-1) and, therefore one obtains, after simple algebraic transformation:

$$\psi^{k+1} = \psi^k + \delta$$

$$\text{div grad } \delta - \delta\cdot(n+p) = n - p - C - \text{div grad } \psi^k + O(\delta^2)$$

$$\text{mit: } n = e^{\psi^k - \phi_n^k}$$

$$p = e^{\phi_p - \psi^k} \qquad (3.1\text{-}3)$$

Because of the exponential nonlinearity of Poisson's equation it is mathematically significant to propose a damping of the potential increment in order to prevent an eventual overshooting of the Newton procedure. This can e.g. be done in the following manner:

$$\psi^{k+1} = \psi^k + \delta/(1+|\delta|/\text{lim}) \qquad (3.1\text{-}4)$$

This form of damping, the so called hyperbolic damping, is continuous with 'lim'·signum (δ) as the boundary value for δ as opposed to ∞ and it has been well proven by all test calculations. 'lim' is in the case of Poisson's equation, when it is represented in its normalized form as in this work, assigned the order of magnitude of 0.25. This way of damping is arbitrary; it has however resulted in a monotonic convergence in all calculations for the Newton procedure for Poisson's equation. The relevant literature offers a wide spectrum of similar variants /14/, /99/.

The linearization of the continuity equation (3.1-2) can be achieved by the application of a Newton-like formulation in a completely analogous way as was briefly explained for Poisson's equation. By suitable substitution of the independent transformation the continuity equation is only very weakly nonlinear, such that, as will become clear in subsection 3.2.3, a linearization can in general be avoided.

3.2 The Discretization of the Fundamental Equations

The differential equations which, in the last section, resulted from the linearization of the complete system of semiconductor equations possess no closed analytic solution and their solution must be found numerically. The first step toward a numerical solution is the partitioning of the regions, in which the differential equations are to be solved under consideration of their boundary conditions, into a finite number of subregions, in which the desired solution to the problem can be approximated with the desired accuracy through simple functions; the equations must be discretized. One must be very careful, because in no case with this procedure does one obtain an exact solution to the analytically formulated problem, instead in the best case an exact solution to the transformed, discrete problem, which, depending upon the fineness of the partioning of the total region and the type of approximate functions in the subregions, represents a more or less good approximate solution to the analytically formulated problem.

There are many classical methods, which propose constructive possibilities for the subdivision of the total region and the choice of approximate functions. In this work a variation of the method of finite differences was used, the five point discretization, which was preferred by most other authors in their work on two dimensional modeling. Examples are the dissertations of Heimeier /62/, of Jesshope /66/ and of Mank /84/ and the review article by Kani /69/ which covered modeling activities in Japan. Discretizations of higher orders were not taken into consideration because of their complexity with respect to programming and possibly for the same reason they also do not appear in the literature.

The method of finite elements is certainly, for purposes of discussion, an alternative with practical relevance which has been used with success by many well known authors (/1/, /8/, /15/, /16/, /30/), and always will be. In the modeling of planar MOS transistors the method of finite differences might be given preference because of the simplicity of the region in which the semiconductor equations are to be solved, and on the grounds of mathematical and physical considerations necessary for partioning of this region. A fundamental mathematical preference for one method or the other is certainly not to be given, therefore in the end the choice is philosophical.

A very interesting modification in certain respects complementing the finite differences method was published by Adler /2/, /3/. With the help of this modification much greater flexibility is achieved in the formulation of the finite differences, which can be very attractive.

In the classical method of finite differences the region in which the solution to the differential equation will be sought is subdivided into subregions through a system of lines parallel to the coordinate axes. Further discussion will be restricted to a cartesian coordinate system and a rectangular solution space, which configuration is exclusively found in the work presented here.

One therefore lays NX vertical (parallel to the y-axis) lines and NY horizontal (parallel to the x-axis) lines through the rectangular region, so one has NX·NY intersection points of these lines, on which an approximate solution for the differential equations is sought. One substitutes only the differential equation on each inner point (i,j) (see figure 3.2-1) through a difference equation, in which the inner point (i,j) - there are exactly (NX-2)·(NY-2) inner points - connected with its four nearest neighboring points (i+1,j), (i,j+1), (i-1,j) and (i,j-1) under the assumption, that the solution of the problem acts as a linear function in the interval spanning these four points and the inner point.

From the already mentioned assumption it becomes clear in an impressive way, how difficult can be the choice for the number of grid lines and their positions for a specific problem. Numerical mathematics has established a considerable amount of evaluation and theories for the purpose of making this choice, e.g. /47/, /140/; in practical cases one needs additionally an enormous amount of experience, in order to advantageously interpret their meaning for a concrete problem.

On the boundary of the region the solution naturally must satisfy the respective boundary conditions, which will be demonstrated later, likewise linear equations can be deduced for the boundary points /47/ - there are exactly 2·(NX+NY-2) points.

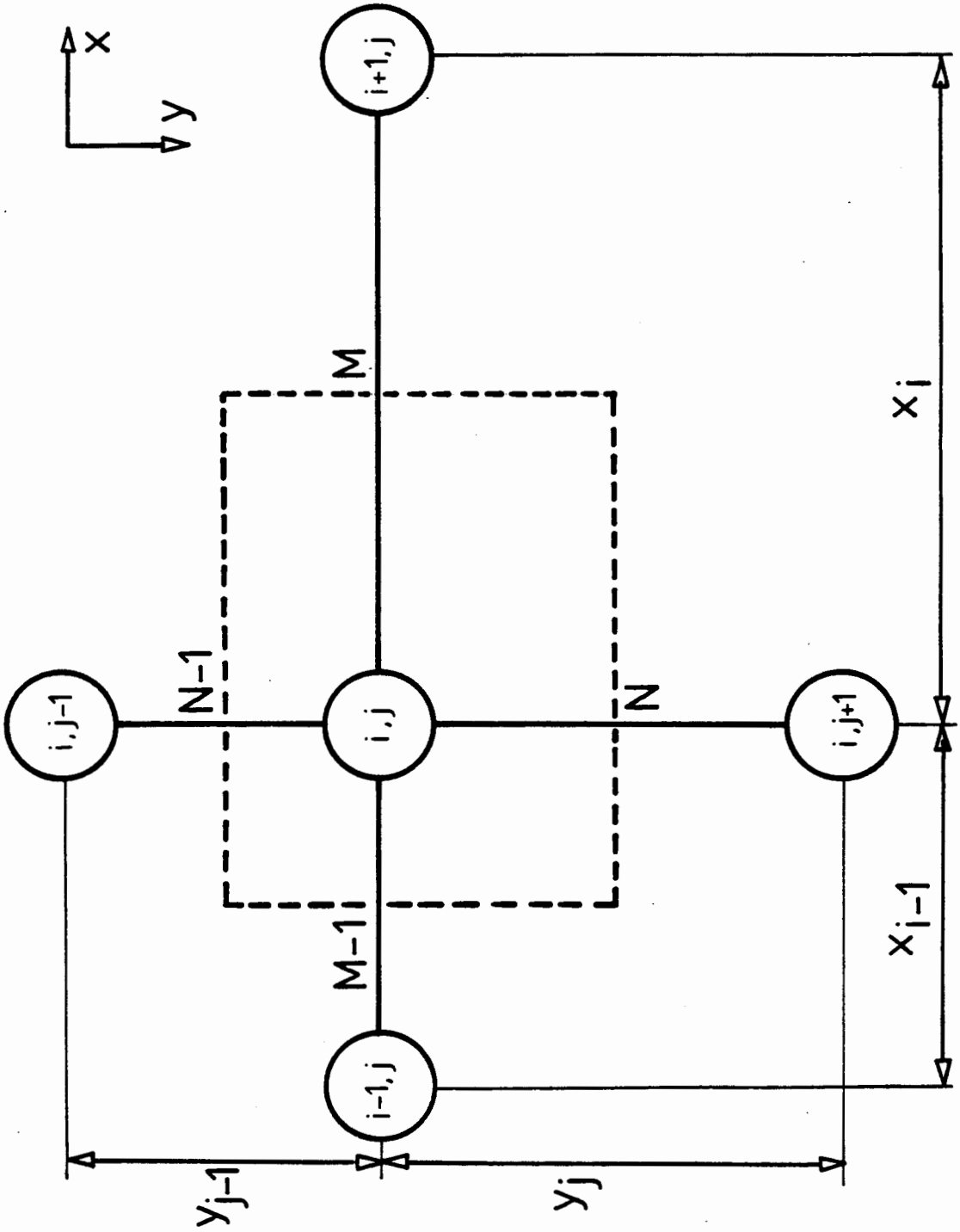


Figure 3.2-1: The adopted nomenclature.

The next subsection will deal with the discretization of the quasiharmonic differential equation, a problem upon which a wide spectrum of literature has been published (e.g. in /47/, /49/, /87/, /120/) and as well the linearized Poisson's equation can, as can also the continuity equation, be transformed to this type by a simple method, which will be shown in the following subsection.

3.2.1 The Quasiharmonic Differential Equation

In the (x,y)-plane there exists a finite continuous region G, which is bounded by a piecewise continuous differentiable boundary R. The functions P(x,y), S(x,y and F(x,y) are continuous in the region G and piecewise continuous on R, the boundary of G. Furthermore P(x,y) is positive and nonvanishing in the complete definition space, likewise S(x,y) and F(x,y) are positive or zero. What is wanted is the function u(x,y) which satisfies the quasiharmonic differential equation:

$$\mathrm{div}(P(x,y)\cdot\mathrm{grad}(u(x,y))) - S(x,y)\cdot u(x,y) = F(x,y) \qquad (3.2\text{-}1)$$

and indeed under the boundary conditions

$$A(x,y)\cdot u(x,y) + B(x,y)\cdot u(x,y)_n = C(x,y) \qquad (3.2\text{-}2)$$

with A(x,y), B(x,y) and C(x,y) defined on R, piecewise continuous and positive or zero, likewise A(x,y) + B(x,y) is positive nonvanishing. $u(x,y)_n$ stands for the externally directed normal derivative of u(x,y) on the boundary R of the region G.

To find the solution to this problem the differential equation will only be integrated in the subregion g_{ij} around the inner point (i,j). This subregion is the rectangle drawn as a dashed line around the point (i,j) in figure 3.2-1.

$$\iint_{g_{ij}} \mathrm{div}(P\cdot\mathrm{grad}(u))\cdot dx\cdot dy - \iint_{g_{ij}} S\cdot u\cdot dx\cdot dy = \iint_{g_{ij}} F\cdot dx\cdot dy \qquad (3.2\text{-}3)$$

With the help of a Green's-like theorem the first surface integral in the above expression can be transformed into a line integral over the boundary r_{ij} of the subregion g_{ij}.

$$\iint_{g_{ij}} \mathrm{div}(P\cdot\mathrm{grad}(u))\cdot dx\cdot dy = \oint_{r_{ij}} (P\cdot(\partial u/\partial x)\cdot dy - P\cdot(\partial u/\partial y)\cdot dx) \qquad (3.2\text{-}4)$$

x_i is the geometrical distance between the i^{th} and $i + 1^{th}$ vertical grid lines and y_j is the distance between the j^{th} and $j + 1^{th}$ horizontal grid lines (see figure 3.2-1). Furthermore P_N stands for the value of the function $P(x,y)$ at the point M, which lies exactly half way between the points (i,j) and (i+ 1, j), and analogously for P_{M-1}, P_N and P_{N-1}, which one can best visualize with the help of figure 3.2-1. It follows that:

$$\begin{aligned}
&\oint_{r_{ij}} (P\cdot(\partial u/\partial x)\cdot dy - P\cdot(\partial u/\partial y)\cdot dx) = \\
&= 0.5\cdot(y_j+y_{j-1})\cdot(P_M\cdot(u_{i+1,j}-u_{i,j})/x_i + \\
&+ P_{M-1}\cdot(u_{i-1,j}-u_{i,j})/x_{i-1}) + \\
&+ 0.5\cdot(x_i+x_{i-1})\cdot(P_N\cdot(u_{i,j+1}-u_{i,j})/y_j + \\
&+ P_{N-1}\cdot(u_{i,j-1}-u_{i,j})/y_{j-1}) + \\
&+ o(x_{i-1}+x_i) + o(y_{j-1}+y_j) \qquad (3.2\text{-}5)
\end{aligned}$$

The second and third surface integrals of (3.2-3) can, under the assumption that the functions $S(x,y)$ and $F(x,y)$ as well as the solution $u(x,y)$ are smooth in the subregion g_{ij}, be integrated in an elementary way.

$$\iint_{g_{ij}} s \cdot u \cdot dx \cdot dy \triangleq 0.25 \cdot s_{i,j} \cdot u_{i,j} \cdot (x_i + x_{i-1}) \cdot (y_j + y_{j-1}) \tag{3.2-6}$$

$$\iint_{g_{ij}} F \cdot dx \cdot dy \triangleq 0.25 \cdot F_{ij} \cdot (x_i + x_{i-1}) \cdot (y_j + y_{j-1}) \tag{3.2-7}$$

One combines (3.2-5), (3.2-6) and (3.2-7) and separates the unknowns, such that one obtains for each inner point (i,j) a linear equation with the following form:

$$\begin{aligned}
& u_{i,j} \cdot ((y_j + y_{j-1}) \cdot (P_M / x_i + P_{M-1} / x_{i-1}) + \\
& + (x_i + x_{i-1}) \cdot (P_N / y_j + P_{N-1} / y_{j-1}) + \\
& + 0.5 \cdot s_{i,j} \cdot (x_i + x_{i-1}) \cdot (y_j + y_{j-1})) = \\
& = u_{i+1,j} \cdot ((y_j + y_{j-1}) \cdot P_M / x_i) + \\
& + u_{i-1,j} \cdot ((y_j + y_{j-1}) \cdot P_{M-1} / x_{i-1}) + \\
& + u_{i,j+1} \cdot ((x_i + x_{i-1}) \cdot P_N / y_j) + \\
& + u_{i,j-1} \cdot ((x_i + x_{i-1}) \cdot P_{N-1} / y_{j-1}) - \\
& - 0.5 \cdot F_{i,j} \cdot (x_i + x_{i-1}) \cdot (y_j + y_{j-1})
\end{aligned} \tag{3.2-8}$$

No residual term for the estimation of the error is provided in (3.2-8). With a nonequidistant grid ($x_i \neq x_{i-1}$, $y_j \neq y_{j-1}$) the error, in the first approximation, is reduced linearly with the line spacing. An exact formulation of the estimated error must be made, however, such a formulation is not a central point of this work but can be found in the relevant literature /47/, /120/.

The discretization of the boundary points presents no problem. It has been carried out in great detail in countless textbooks on numerical mathematics (e.g. /47/), therefore it is refrained from a simple repetition.

One combines the equations for all of the grid points, which are linear without exception, in one system, such that this system can be presented in matrix form.

$$A \cdot u = b$$

In the above equation u represents a vector of length NX NY, in which all of the desired $u_{i,j}$ are included. b is the pertinent vector for the right hand side. The matrix A in many practical cases is of very high rank NX NY. As the coefficient matrix of the system equations it has a maximum of five non zero elements in each row. The numerical solution of a system of sparsely linear equations of this type will be dealt with in section 3.3.

3.2.2 Poisson's Equation

The linearized form of Poisson's equation (3.1-3) is tailor made for the quasiharmonic equation whose theory was sketched in the last subsection. In order to see this established in its entirety, it is necessary to clarify the boundary conditions of Poisson's equation as well as the structure of the equation to be solved, and also the appropriate boundary conditions are necessary in order to consider and thereby classify the problem clearly and place it in proper perspective.

Figure 3.2-2 shows the geometry which was used in the simulation. Inside the rectangular region A-F-G-H, which represents the silicon, the system of semiconductor equations must be solved, in the region C-D-E-B which represents the gate oxide, only Laplace's equation must be solved.

The following concrete considerations apply only to Poisson's equation. The contacts (AB: Source, EF: Drain, GH: Bulk) will be assumed to be ohmic. The potential will hereafter have the value of the applied potential plus the built in potential due to the doping. On the vertical edges of the semiconductor (AH, FG) the normal derivative of the potential, that is, the lateral electric field component, must go to zero. This consideration is naturally justified when the vertical edges are far enough away from the channel. On the interface (line BE) equation (3.2-9) must satisfy Gauss's law.

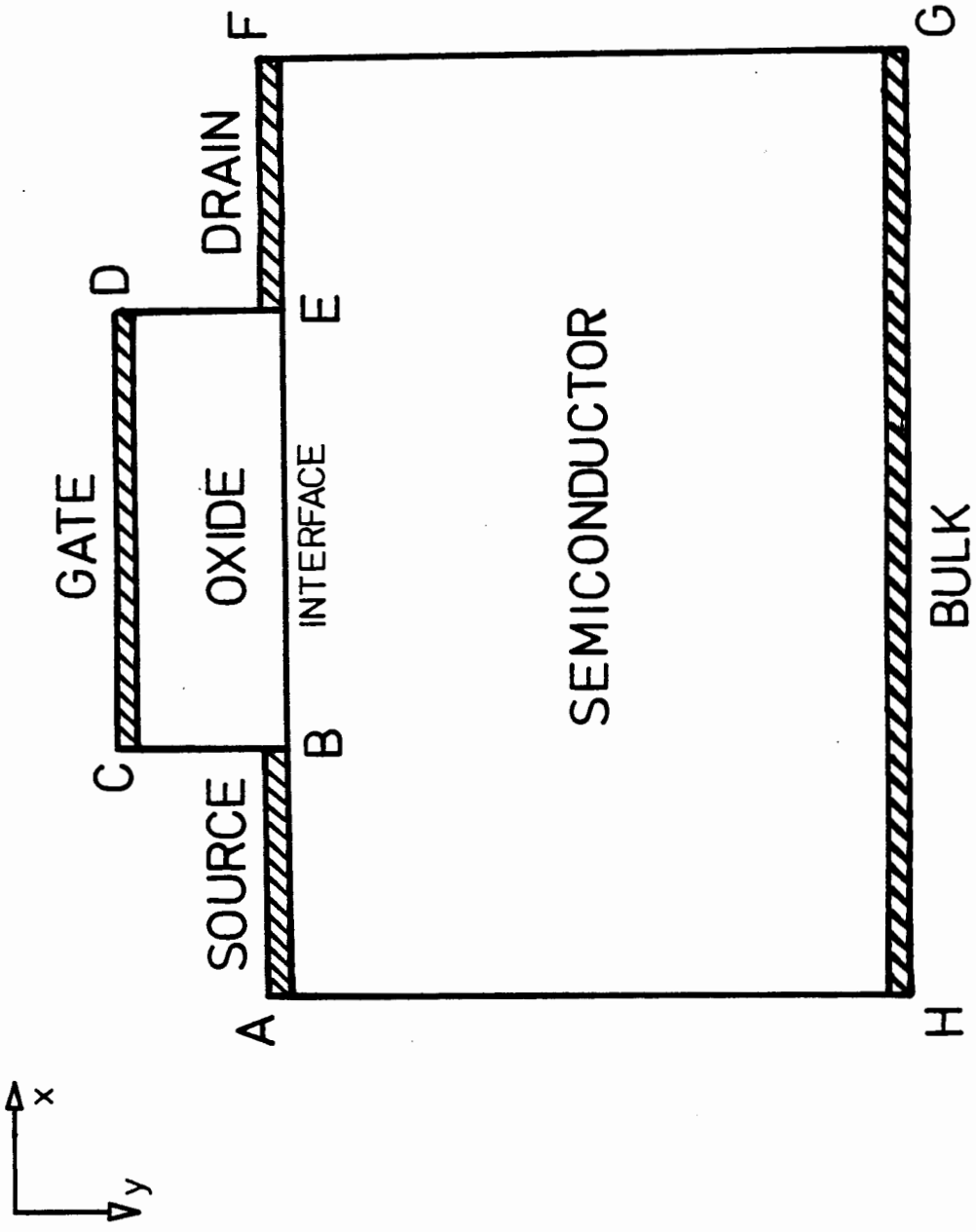


Figure 3.2-2: The simulation geometry.

$$\varepsilon_{ox} \cdot (\partial\psi/\partial y)_{ox} = \varepsilon_{si} \cdot (\partial\psi/\partial y)_{si} \qquad (3.2\text{-}9)$$

In the oxide a solution is to be found for Laplace's equation which is coupled to Poisson's equation in the semiconductor, consequently (3.2-9) represents the boundary condition which must be satisfied along the interface. On the vertical edges of the oxide (CB,DE) the normal derivative of the potential must again go to zero, and on the gate contact (CD) the potential will be set equal to the applied voltage minus the flat band voltage.

The eventual existence of slow interface states N_{ss} will be taken into account in the flat band voltage. Gauss's law permits a physically consistent consideration of the accumulation of charge by the slow interface states /126/. Due to the decreased importance of these interface states in modern transistors, the added expense of this alternative is not justified.

With the above defined boundary conditions it is easy to verify that the linearized form of Poisson's equation is a special case of the boundary value problem dealt with in 3.2.1. Only the following substitutions need to be made:

$u(x,y)$ ... $\psi(x,y)$

$P(x,y)$ ... 1

$S(x,y)$ ... $n(x,y) + p(x,y)$

($n(x,y)$ and $p(x,y)$ contain only the known values of ψ^k and φ_n^k)

$F(x,y)$ ... $n(x,y) - p(x,y) - C(x,y) - \text{div grad } \psi^k$

For the boundary one has:

in the cases of the contacts

$A(x,y) = 1$

$C(x,y) =$ const

(applied plus built-in potential)

$B(x,y) = 0$

in the cases of the vertical boundaries

$A(x,y) = 0$

$B(x,y) = 1$

$C(x,y) = 0$

in the case of the interface:

$A(x,y) = 0$

$B(x,y) = \varepsilon_{si}$

$C(x,y) = \varepsilon_{ox} \cdot (\partial\phi/\partial y)_{ox}$

Laplace's equation in the oxide is only a trivial case of Poisson's equation and will not be given further consideration here.

The discrete form of Poisson's equation can, with the help of the substitution presented in (3.2-8), be determined in an elementary way, by way of a direct copy of (3.2-8) with a partial change of notation.

3.2.3 The Continuity Equation

At first glance the continuity equation does not fit the theory of the quasiharmonic differential equation. First by the transformation of variables

$$s = e^{-\phi n} \qquad (3.2\text{-}10)$$

the required analogy will become evident. The continuity equation has the following form:

$$\begin{aligned} &\mathrm{div}(\mu_n \cdot e^{\phi} \cdot \mathrm{grad}\, s) - s \cdot e^{\phi} P/(\tau_p \cdot (n+1) + \tau_n \cdot (p+1)) = \\ &= -1/(\tau_p \cdot (n+1) + \tau_n \cdot (p+1)) \end{aligned} \qquad (3.2\text{-}11)$$

The linearization of the denominators of the recombination terms is not necessary here, in the case of MOS transistors, because on the one hand this term makes no conbribution to the basic carrier transport, but serves only to avoid physically unrealistic carrier depletion in the drain/substrate diode (see chapter 2) and on the other hand, through the use of an iterative solution procedure for the differential equation (see section 3.3) a first order linearization occurs automatically.

The boundary conditions are simple as was the case for Poisson's equation. The contacts (AB, EF and GH in figure 3.2.2) are taken as ohmic,

and the carrier concentrations are equal to their equilibrium values. On the remaining boundaries (BE, FG and AH) no normal current components may exist at any time.

The reader can easily see, by way of the above considerations, that the continuity equation fits into the scheme of boundary value problems dealt with in section 3.2.1, such that further discussion in this direction can be dispensed with.

The above presented derivation may indeed work very satisfactorily and completely, but still there are several additional tricks to be applied, without which the transformation of this theory into a computer program would surely be condemned to failure.

The first problem that exists is to choose a suitable interpolation of the function in (3.2-12):

$$P(x,y) = \mu_n \cdot e^{\psi} \qquad (3.2\text{-}12)$$

The mobility is indeed, in general, a function with a small variation, such that it can be linearly interpolated between neighboring grid points. The exponential function of the potential surely cannot be interpolated in the same way. A simple method is the geometrical averaging of the function, which agrees with a linear behavior of the potential. An error analysis of this method has been carried out by Jesshope /66/, /67/. This interpolation will not be exact even when the potential difference between two adjacent points is small. A satisfactory interpolation algorithm was first published by Scharfetter and Gummel /110/. That algorithm was in essence based upon physical considerations; its mathematical consistency was verified by Barnes /7/. It holds that for the interpolation between the points (i,j) and (i + 1, j):

$$\begin{aligned} e^{\psi}_{M} &= e^{\psi_{i,j}} \cdot (\psi_{i,j} - \psi_{i+1,j}) / (e^{\psi_{i,j} - \psi_{i+1,j}} - 1) \\ &= e^{\psi_{i,j}} \cdot \mathrm{ber}(\psi_{i,j} - \psi_{i+1,j}) \end{aligned} \qquad (3.2\text{-}13)$$

with: $\mathrm{ber}(x) = x/(e^x - 1)$ (The Bernoulli function).

Special attention must be paid to the programming of the Bernoulli function in order to avoid overflow or underflow of the numerical range of the computer /60/. A fortuitous secondary effect of using this interpolation is, that, when one divides the therewith obtained difference equation by $e^{\psi i,j}$, which causes absolutely no problem, only exponential functions of potential differences appear in the coefficients, which decisively increases the numerical stability of the system of equations.

A further point which requires close attention for the solution of the discretized continuity equation is the extremely large interval for S, (3.2-10) which has been obtained by substitution of a new independent variable. It is finally necessary to reduce the solution interval with the help of a similarity transformation. A physically based scaling rests upon the use of e^{ψ} as the transformation variable /117/. From this method one obtains as independent variable the carrier concentration, whose dynamic range can indeed be very large but presents no difficulty for modern computers. An unfortunate side effect of this similarity transformation is that the symmetry of the coefficient matrix is lost. The global condition number of the system of equations will not be changed through the similarity transformation /39/.

An interesting alternative to the discretization of the continuity equation which was sketched here is to mention the connection with the "stream function" method which was first published by Mock /90/. The stream function method was not tested in this work, the method presented above functioned satisfactorily. The method of "stream functions" enjoys great popularity especially with the Japanese authors /69/, /131/. An evaluation of this method on purely theoretical considerations is very problematic and therefore will not be attempted here.

3.2.4 The Grid Generation

Careful attention must be given to the selection of the grid points in order to sufficiently bound the discretization error, which has a very strong influence upon the convergence characteristics of the equations and therewith upon the complete system.

It is impossible to choose an equidistant grid, because the spacing must be compatible with the dominant region of greatest numerical difficulty. Consequently, the use of an equidistant grid would result in too many unnecessary points in other regions. Therefore an enormously large number of required grid points would result, which would cause noticeably undesirable memory and computation time requirements. A further, more severe disadvantage also exists. In a region in which a solution variable is almost constant, small spacings can result in numerical instabilities which can be traced to the round off error which results from taking the difference between two nearly equal numbers.

Because of the above mentioned reasons, a grid with unequal spacing must be used, which will then be checked for accuracy in a special phase of the solution (see appendix C). It will then, if necessary, be adapted where the most recently obtained behavior of the variable will be used as the new basis for the calculation.

The underlying conditions for the grid generation can be divided into two different groups: there is fundamentally, on the one hand, the spacings in regard to these conditions, which are to be fulfilled under all circumstances, and on the other hand the conditions are considered on the grounds of doping profile, electrostatic potential and carrier distribution, so far as these last conditions do not conflict with the first group. It should possibly be mentioned here that the grid generations in the x and y directions are independent and therefore no conditions exist in this regard.

Requirements which take absolute priority are that there exists a minimum value for the spacing and that the ratio of successive spacings must be between certain minimum and maximum values, whereby the maximum allowable progression of the grid is established.

The considerations with respect to the doping profiles exist as criteria relative to the active doping concentration between two successive grid points, which is only important when at least one of the two doping levels is greater than some minimum value.

Respectively the electrostatic potential must satisfy the requirement that the absolute potential difference between two adjacent grid points may not exceed some maximum value.

A requirement imposed by the charge carrier concentration is that the ratio of the electron density (or the hole density for p-channels) between two adjacent points must lie between certain minimum and maximum values.

The setting of the bounds for each of the above requirements by strong mathematical considerations is highly unrealistic, because some remainder may arise which can only be evaluated by an enormous, unjustified expense. The values of the bounds which were actually used came from intuitive or physical considerations and were proven plausible by many test cases. The criteria were placed in a separate part of the program in a modular fashion such that changes could be easily made.

3.3 The Solution of the Discretized Fundamental Equations

Given the system of equations

$$A \cdot u = b$$

with the property that the coefficient matrix A is derived from a five point discretization using finite differences. In general the rank of this matrix is very large (typically 2000-3000) and furthermore there are only a maximum of five elements in each row and column which are not equal to zero. Consequently, the matrix is very sparsely filled, it is sparse. Because of the discertization of Poisson's equation the resulting matrix is symmetrical and positive definite. In principle these properties also hold for the coefficient matrix of the continuity equation, whereby it should be noted, that these properties are lost through an eventually required similarity transformation. Further properties of these matrices will not be required for the following discussion, therefore, the relevant literature can be referenced /87/, /139/, and /140/.

Two kinds of approaches can be used for the solution of these types of linear systems of equations: direct and iterative approaches. The classical direct method, Gaussian elimination, does not take into account the sparseness and the special structure of the coefficient matrix, such that the required computational effort (rank $(A)^3$) is in general, not acceptable. However, there exists a remarkable number of modified Gaussian elimination methods /43/, which consider to some extent the above mentioned properties of the coefficient matrix. Because the solution of the system of equations is embedded in an iteration procedure, it therefore very often must occur that priority has to be given to the iterative methods for the solution of the system of equations. This holds even more as through the iteration procedure very good initial guesses are available /42/.

In the scope of this work a large number of iterative solution procedures were programmed and tested. The relaxation processes (SOR, LSOR, SSOR, S2LOR) were ruled out in the first test comparison because of their slow rates of convergence. The programming of these methods is indeed very easy and there

exists a wide spectrum of literature /122/, /139/, and /140/ in which their convergence characteristics and their mathematical foundations are exactly analysed, such that the attraction of these methods is that they are very serviceable. After careful consideration the choices were limited to the ADI method /122/, the AFP method /44/ and the SIP method of Stone /123/. Careful comparison of these three methods indicated an unmistakable advantage for the SIP methods, as was also confirmed in /105/.

The basic idea of Stone's method is that a special matrix N is added to the coefficient matrix A, such that the resultant matrix (A + N) can be decomposed trivially into the product of an upper triangular matrix and a lower triangular matrix.

$$(A + N)\cdot u = (L\cdot U)\cdot u$$

Under this condition the construction of an iterative process is simple. It holds namely:

$$(A + N)\cdot u = (A + N)\cdot u = (b - A\cdot u)$$

hence one can obtain:

$$(A + N)\cdot u^{k+1} = (A + N)\cdot u^{k} + (b - A\cdot u^{k}).$$

Because the right side of this system of equations is known and because (A + N) is simply factorizable, the above equation represents an extremely efficient iteration scheme. Furthermore, if the norm of N is very much smaller than the norm of A, a fast convergence rate can be intuitively expected.

Stone further gave a simple, and constructive possibility for the choice of the matrix N; closer examination regarding this topic would greatly expand the scope of the present work, therefore those interested are referred to the original work /123/ or the explanations of this procedure in /49/ and /120/.

An eventual disadvantage of this method is the fact, that the vectorization of this algorithm in view of a computer with pipeline architecture, as is also noted in /42/, and the efficient use of fundamental modular programs of linear algebra /77/ are not simple.

4. Typical Applications Examples

The demonstration of a finished work by typical examples is unquestionably one of the most important points of the work itself. Examples first bring life to abstract formulas; they stimulate the imagination, induce ideas and thereby often times build a graphic basis for further and analogous works. The appropriate choice of such examples can be very difficult when many such possibilities exist, one must appeal to a wide spread public yet only a small space is available.

Three significant examples will be presented in this work. Subsection 4.1 represents a didactic example. On the one hand this example should be easily understandable and only for the general interest in the simulation without specific knowledge of the MOS device and on the other hand it should provide interest and stimulation for the experienced reader. In subsection 4.2 the simulation of an inverter will be dealt with. This example should appeal to the designer of circuits with miniaturized devices and also to the device designer. In the third subsection (4.3) the problem of the process sensitivity of modern transistors, which is one of the general interests of technology at the present, will be examined.

Extremely high quality graphics are presented in all three subsections. A multiple number of figures, which show the physical distributions of the relevant quantities within a greatly enlarged cross-section of the interior of the transistor will be used. Because of this, subtle details can frequently be explained. In the examples the comparisons of calculated and measured characteristic curves were omitted. Comparisons of this type would of course provide increased verification of the numerical models and of the limitations of the tolerances of the physical parameters of the models. In general, good agreement between calculated and measured characteristics curves can be obtained. An important consideration here is that one must know, with good accuracy, the relevant technological parameters and geometry of the transistor which is measured. Only then, because of the increased process sensitivity of miniature transistors, is there a chance to achieve the desired results. For the purpose of testing and verification, MINIMOS has been passed on, by way of academic exchange, to a large number of highly interested international semi-

conductor manufacturers. A realistic verification and evaluation with constructive criticism is surely only fair and serious with the aid of many users and a wide spectrum of transistors from different manufacturers.

It is to be further mentioned, that no transistors were fabricated in the scope of this work. The availability of transistor material from outside firms was relied upon completely. A proper fabrication of transistors was, by the best intentions, not possible because the necessary facilities did not exist.

4.1 A Didactic Example

In general, it is difficult to present an application example of two dimensional modeling which on the one hand is interesting to the experienced reader and on the other hand is easily understood by those who have a general interest in this work but who have no specific knowledge of the MOS system. A thoroughly appropriate example of this type is the analysis of the influence of an ion implantation upon short-channel transistors. For that purpose three transistors were simulated whose data are declared in the MINIMOS input statements (Figure 4.1-1). At this point it should also be mentioned, that the transistors discussed in the following were never actually fabricated; rather their data were chosen so as to demonstrate the distinct effects of the analyses. They could, however, be directly realized in any good laboratory, because the specified data are technologically significant. As will become clear in the following, the third transistor is thoroughly suited for use in an integrated switching circuit with a one micrometer technology.

The first line of each of the three input statement sets (Figure 4.1-1) is a title, which will identify the computer output; it is therefore simply for commentary purposes. The further syntax (as detailed and discussed in appendix A) is based upon a keyword-parameter-value-structure and is completely format free.

```
ONE-MICROMETER ANALYSIS (TRANSISTOR 1)
DEVICE      CHANNEL=N  GATE=NPOLY  TOX=350.E-8 W=10.E-4 L=1.E-4
BIAS        UD=3.  UG=0.
PROFILE     NB=1.E15  ELEM=PH  DOSE=1.E15  AKEV=40  TOX=350.E-8
+           TEMP=1000  TIME=1200
END

ONE-MICROMETER ANALYSIS (TRANSISTOR 2)
DEVICE      CHANNEl=N  GATE=NPOLY  TOX=350.E-8 W=10.E-4  L=1.E-4
BIAS        UD=3.  UG=0.
PROFILE     NB=1.E15  ELEM=PH  DOSE=1.E15  AKEV=40  TOX=350.E-8
+           TEMP=1000  TIME=1200
IMPLANT     ELEM=B  DOSE=3.5E11  AKEV=25  TEMP=925  TIME=1800
END

ONE MICROMETER ANALYSIS (TRANSISTOR 3)
DEVICE      CHANNEL=N  GATE=NPOLY  TOX=350.E-8 W=10.E-4 L=1.E-4
BIAS        UD=3.  UG=0.
PROFILE     NB=1.E15  ELEM=PH  DOSE=1.E15  AKEV=40  TOX=350.E-8
+           TEMP=1000  TIME=1200
IMPLANT     ELEM=B  DOSE=3.5E11  AKEV=25  TEMP=925  TIME=1200
IMPLANT     ELEM=B  DOSE=1.5E11  AKEV=100
END
```

Figure 4.1-1: Typical input statement sets.

The second line with "DEVICE" as the key word describes the type and geometry of the transistor. An N-channel device is specified (CHANNEL=N) with an N-type polysilicon gate (GATE=NPOLY) and with an oxide thickness of 35 nanometers (TOX=350.E-8), a channel width of ten micrometers (W=10.E-4) and a channel length of one micrometer (L=1.E-4).

The operating point is established by the "BIAS" input. A drain voltage of three volts (UD=3.) and a gate voltage of zero (UG=0.) were chosen. When a substrate voltage is not given explicity, MINIMOS will assume a value of zero volts.

The substrate doping and the source-drain profiles are specified by the "PROFILE" input. In these examples the simplest means of defining the doping profiles was chosen: a direct calculation with MINIMOS. A substrate doping of 10^{15} cm^{-3} (NB=1.E15) and a source/drain implantation with phosphorous (ELEM=PH), and an implantation dose of 10^{15} cm^{-2} (DOSE=1.E15) and an implantation energy of 40 keV (AKEV=40) were selected. The implantation was made through an oxide with a thickness of 35 nanometers (TOX-350.E-8) and was annealed at 1000 degrees centigrade (TEMP=1000) for 1200 seconds (TIME=1200).

The second set of input statements also contains an "IMPLANT" specification for the channel implantation with Boron (ELEM=B), a dose of $3.5 \cdot 10^{11}$ cm^{-2} (DOSE=3.5E11) and an energy of 25 keV (AKEV=25). The anneal was at 925 degrees centigrade (TEMP=925) and 1800 seconds long (TIME=1800).

The third set of input statements contains a second "IMPLANT" input statement for a second, deeper channel implantation with Boron (ELEM=B), a dose of $1.5 \cdot 10^{11}$ cm^{-2} (DOSE=1.5E11) and an energy of 100 keV (AKEV=100). It is assumed by MINIMOS that both channel implants are annealed together.

It is surely well known by many readers, that the first transistor in these examples will exhibit a small negative threshold voltage due to short channel effects and, that the first shallow channel implantation serves to shift the threshold voltage to a positive value.

The deep channel implantation is for the purpose of eliminating the eventual problem of "punch-through". This effect will be illustrated further by 3-D figures of distributions of the relevant physical quantities in the interiors of the transistors.

Figures 4.1-2, 4.1-3 and 4.1-4 show the doping profiles calculated by MINIMOS.

One can accurately read the source/drain pn-junction depth of 300 nanometers from these figures. The surface concentration in these highly doped source/drain regions is about 10^{20} cm^{-3}. The effective channel length will be reduced by the lateral diffusion of about 0.6 micrometer. The shallow

channel implantations are easy to distinguish in figures 4.1-3 and 4.1-4. The deep channel implantation in the third transistor (Figure 4.1-4) for the suppression of the "punch-through" effect exhibits a local doping maximum at about the same depth as the source/drain pn-junction. The deep implantation will have a negligible influence upon the threshold voltage.

Figures 4.1-5a,b show the distribution of the electrostatic potential in the first transistor. For clarity of presentation Figure 4.1-5b shows a graphical representation in the form of equipotential lines. The drain contact is at the right rear of the 3-D figure. The zero potential point was established at the middle of the forbidden band. In the space charge zone of the reverse biased drain/substrate diode the potential falls off monotonically; and in the highly doped source/drain regions it appears constant. One observes only a very small barrier between the source and the channel. Figures 4.1-6a,b show the potential distribution in the second transistor. In the 3-D representation of the potential only a small difference is observed with respect to the first transistor; the source/channel diode barrier is slightly more distinct here. In the equipotential representation one observes a local potential minimum directly under the surface. The significance of this local potential minimum is that there exists a saddle point under and slightly to the left of the local potential minimum. This saddle point is a field free point in which current can only flow as diffusion current. A saddle point of this type as has been dealt with by many authors (e.g. /9/, /75/) is typical for the occurrence of the "punch through" effect.

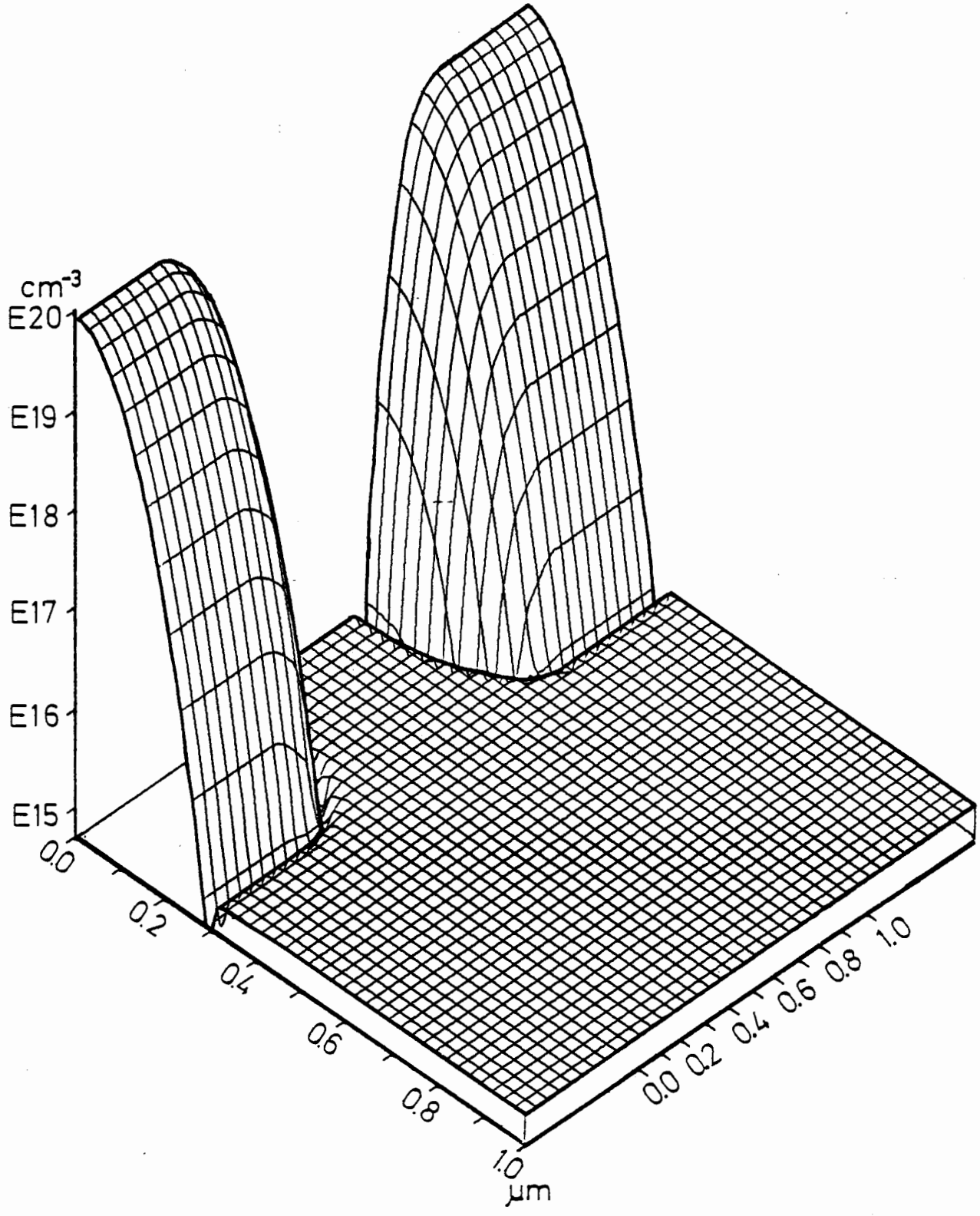


Figure 4.1-2: Doping profile of the first transistor.

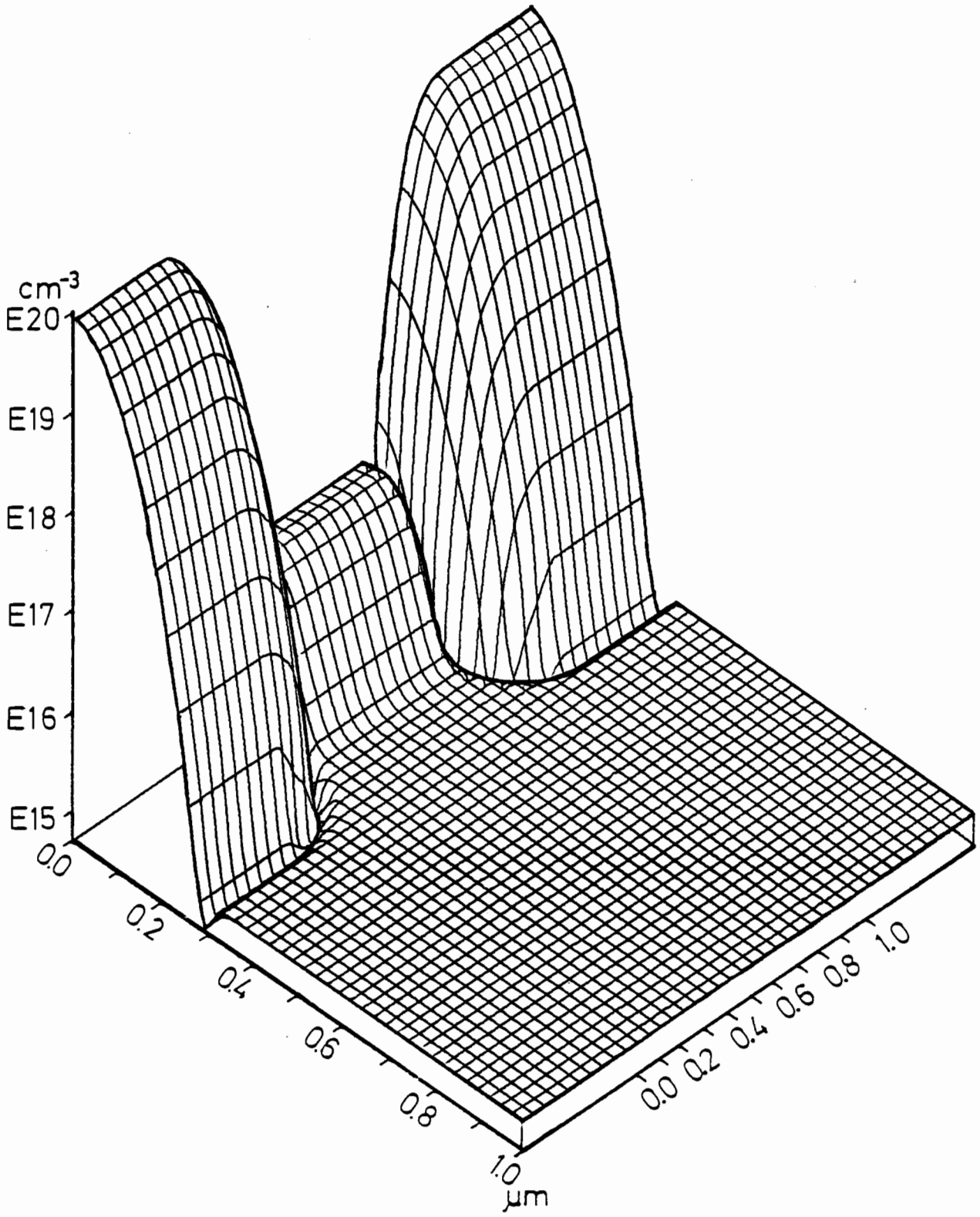


Figure 4.1-3: Doping profile of the second transistor.

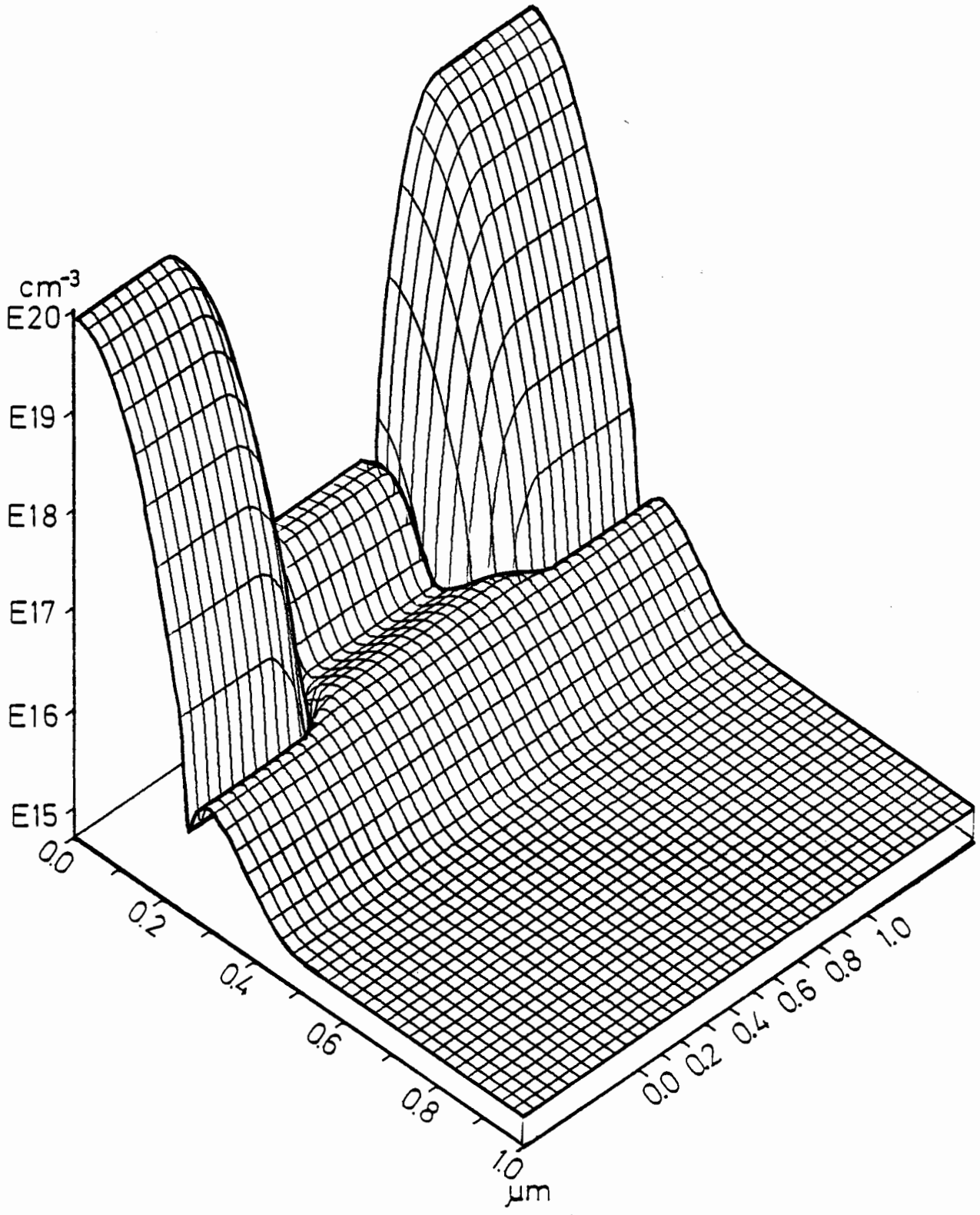


Figure 4.1-4: The doping profile of the third transistor.

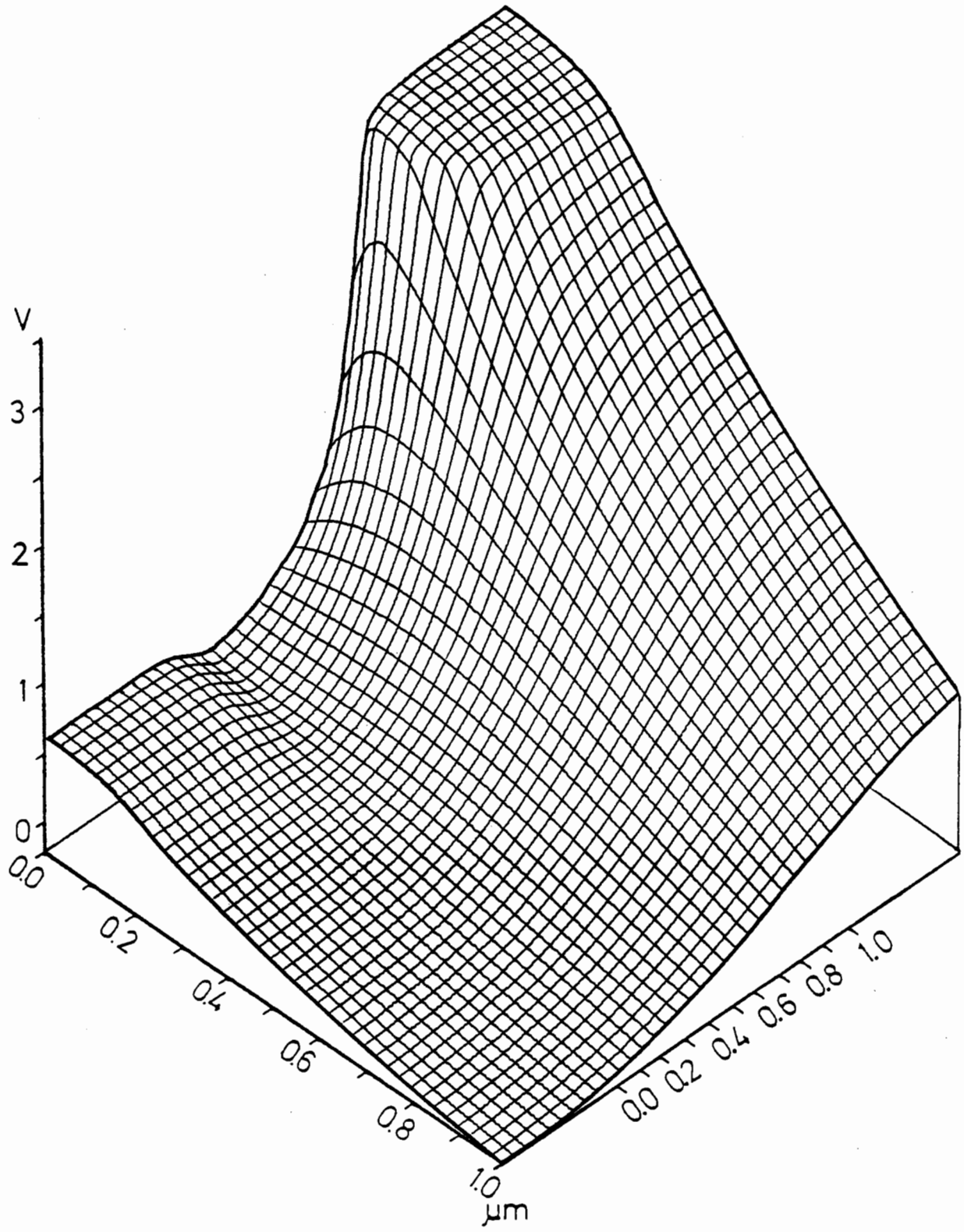


Figure 4.1-5a: Potential distribution in the first transistor.

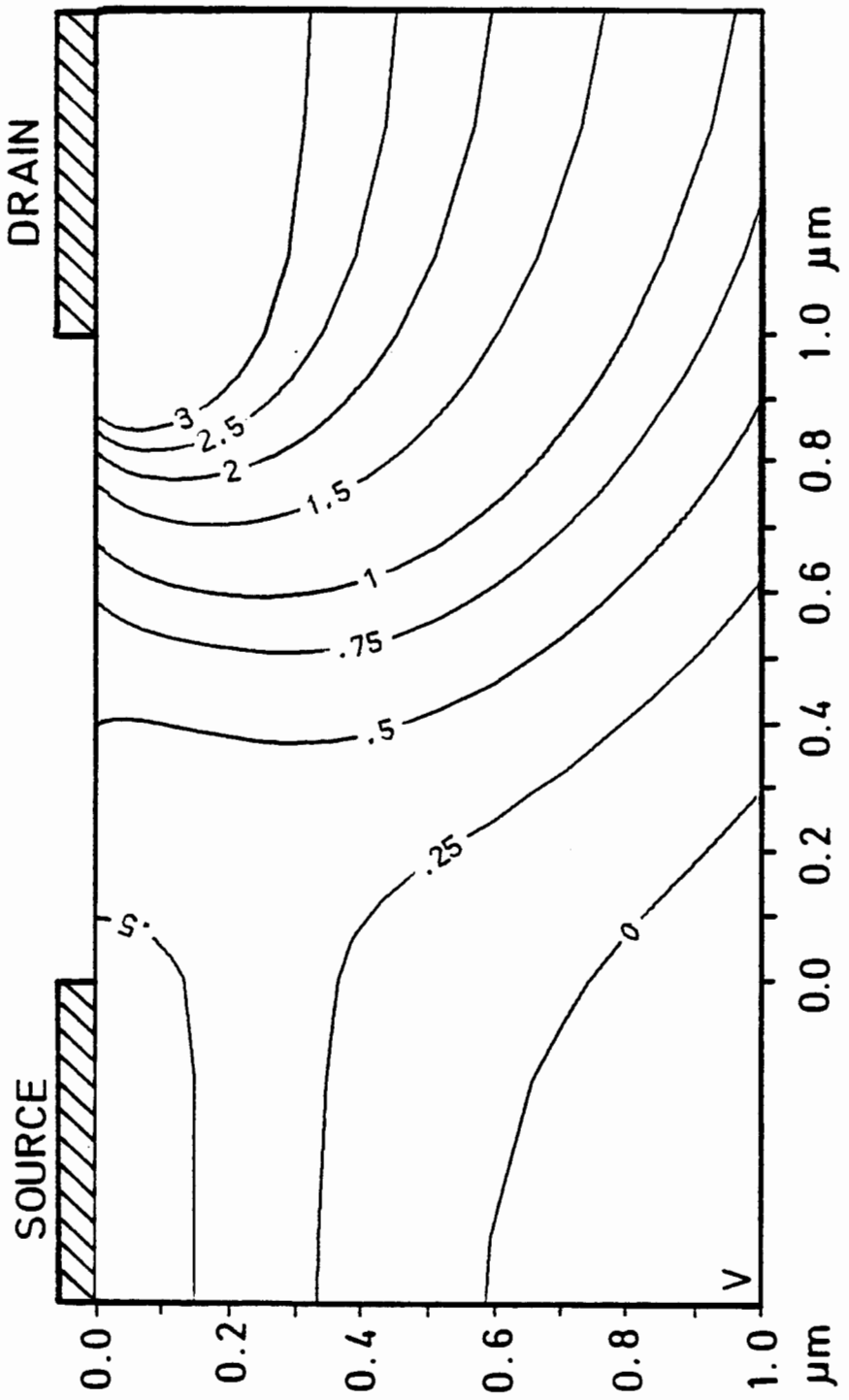


Figure 4.1-5b: Potential distribution in the first transistor.

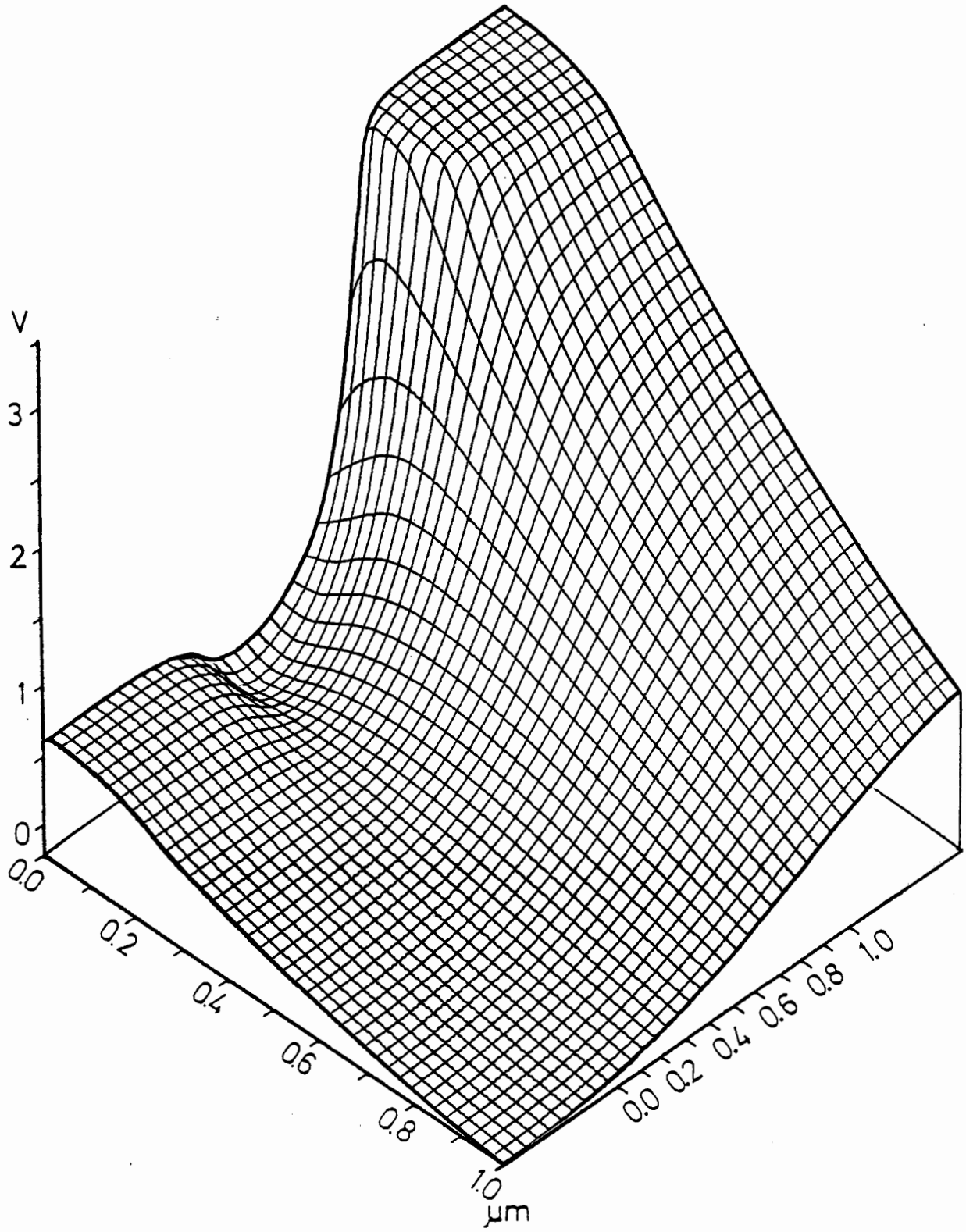


Figure 4.1-6a: Potential distribution in the second transistor.

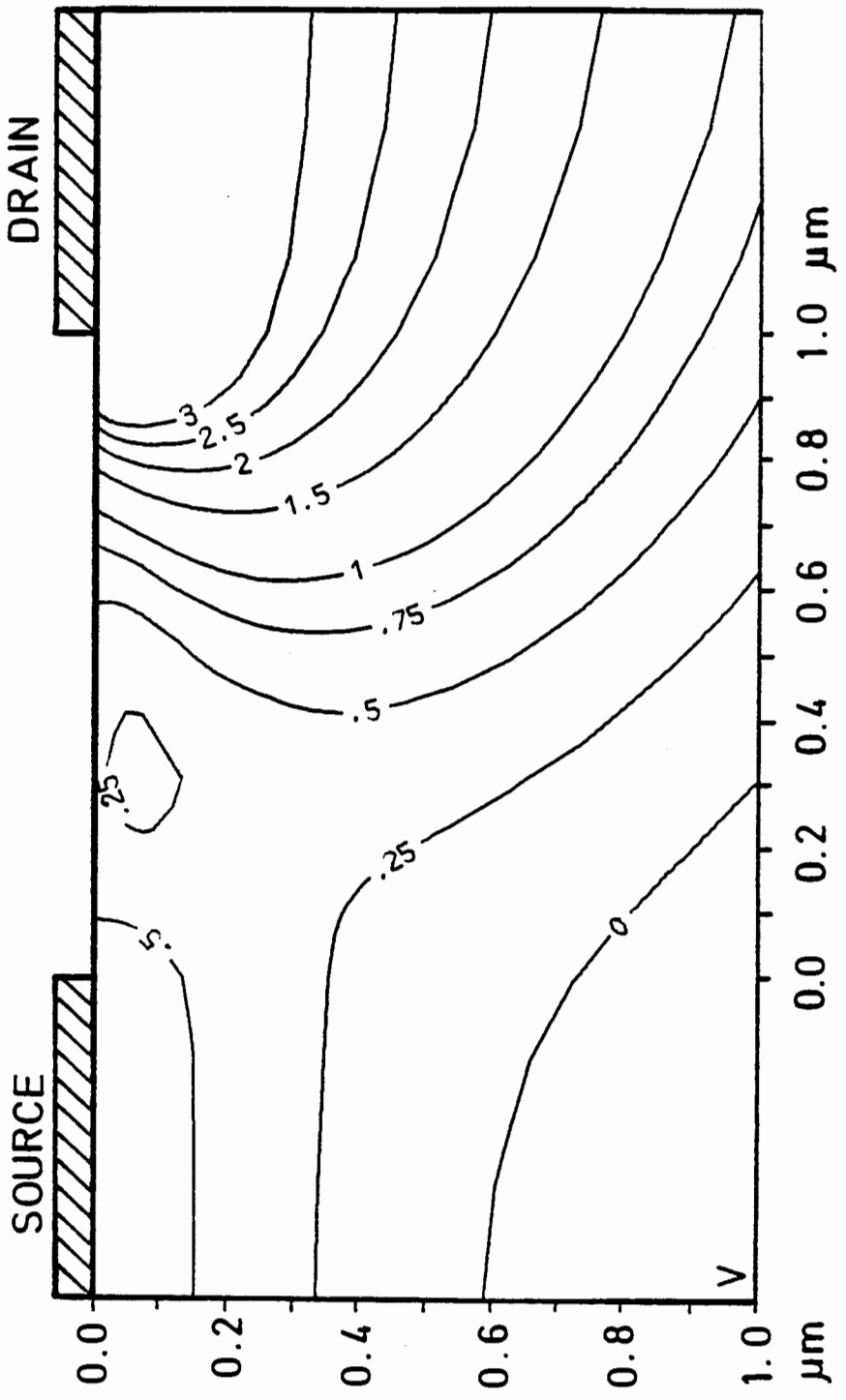


Figure 4.1-6b: Potential distribution in the second transistor.

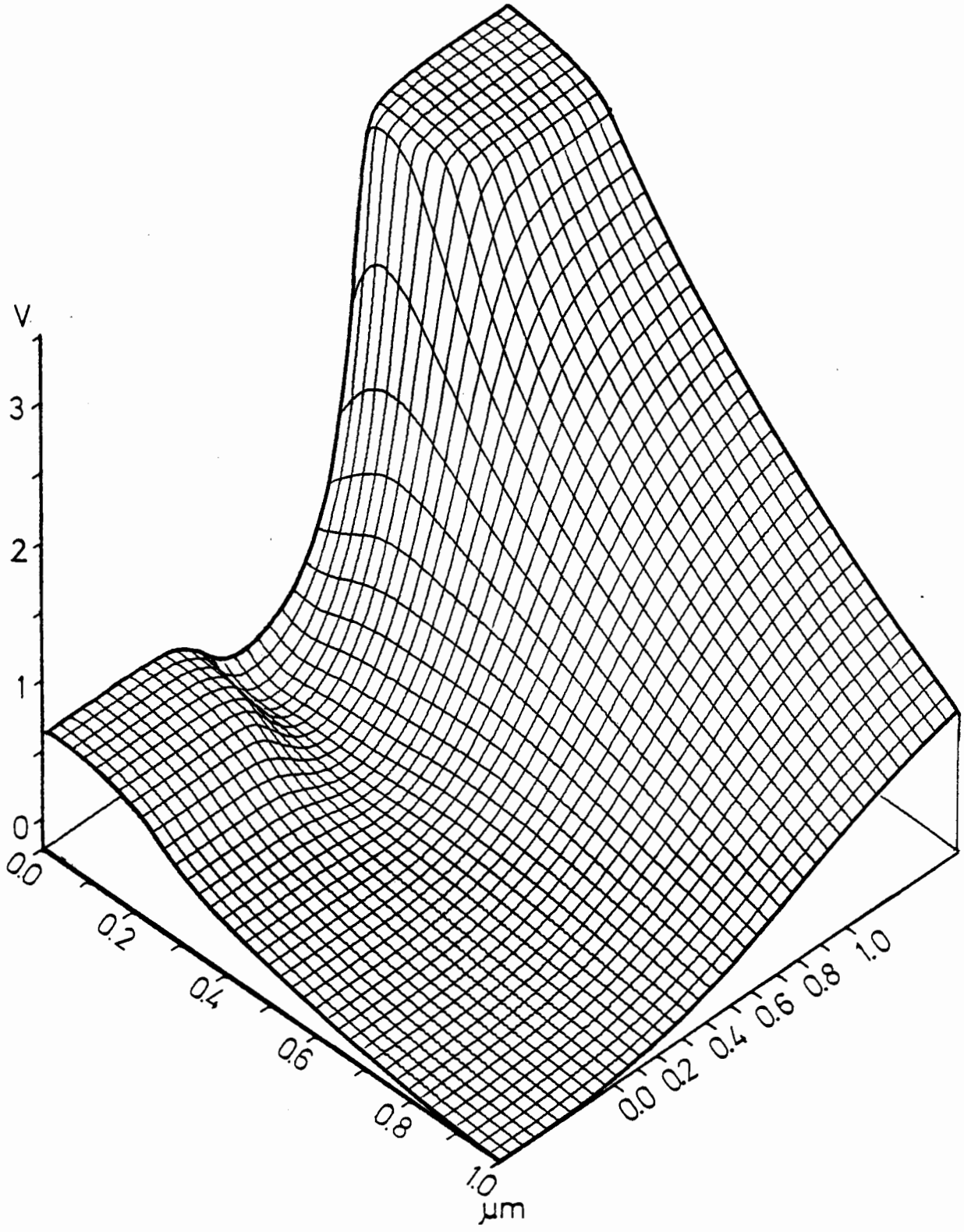


Figure 4.1-7a: Potential distribution in the third transistor.

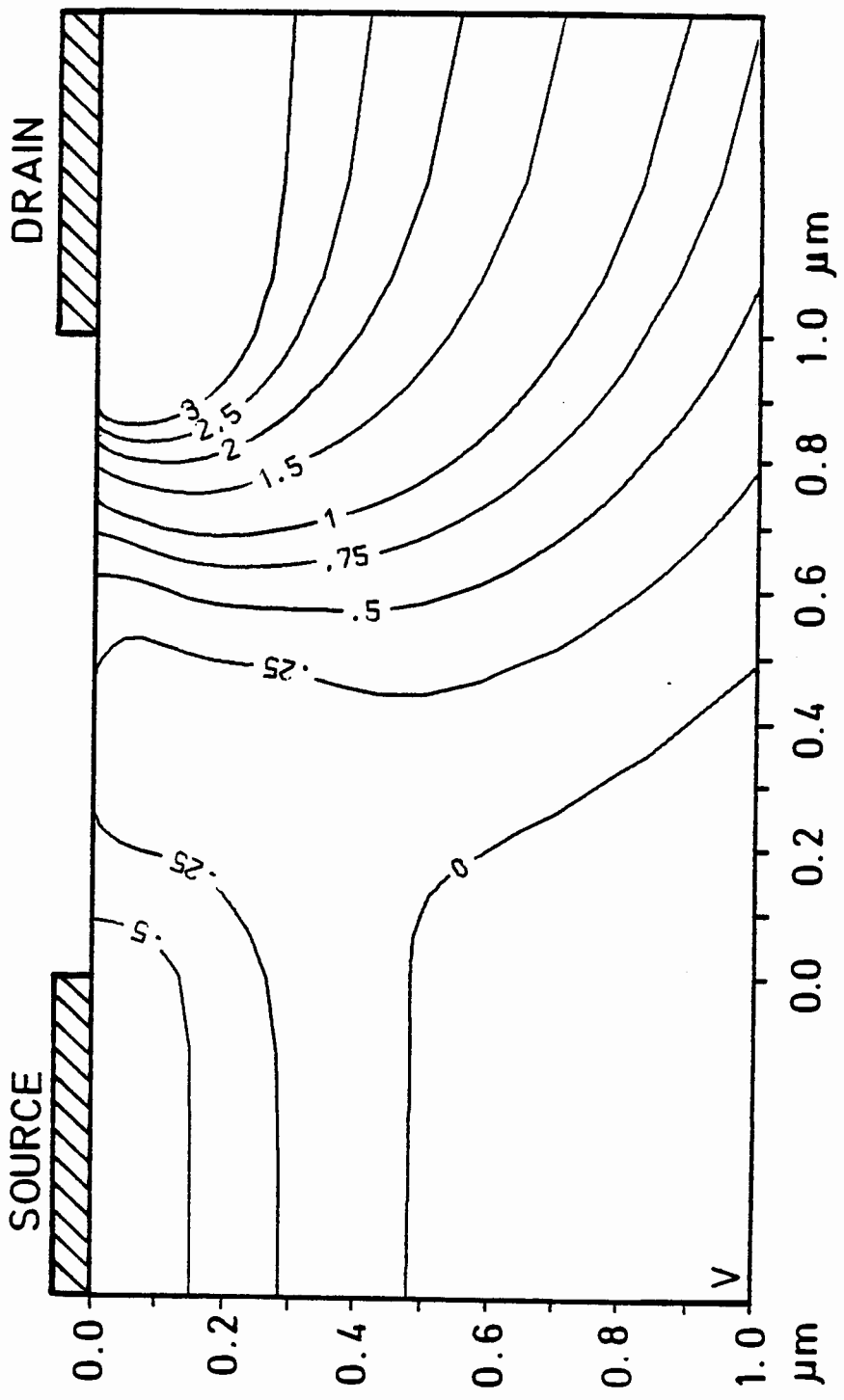


Figure 4.1-7b: Potential distribution in the third transistor.

In cases such as these it is not sufficient to separate the gate induced field from the space charge zones of the source and drain. This effect is partially under the influence region of the gate. As will become more clear later, the occurrence of a potential saddle is a sure indication of "punch through", but it is by no means necessary.

Figures 4.1-7a,b show the potential distribution in the third transistor. The 3-D representation shows only a marginal difference in comparison with the second transistor. From the equipotential graph one can observe a source/channel barrier which is well enough defined so as to guarantee a properly operating transistor at the given operating point.

Figures 4.1-8a,b,c show the electron density distribution of the first transistor. In order to improve the clarity of the presentation all figures which show carrier density distributions include two 3-D representations and a constant density contour plot of the carrier concentration. Figures 4.1-8a, 9a and 10a show the density in the usual orientation with the drain contact at the right rear. Figures 4.1-8b, 9b and 10b show the density likewise at the surface of the transistor. This makes it possible to view the channel region from the surface.

The surface concentration of electrons in the channel is relatively high, which is due to the small negative threshold voltage. The operating point lies in the region of strong inversion. In the region of the drain contact a very distinct depression in the surface carrier concentration is observed, which represents the so called "pinch off" zone. It is also shown that the region of high electron density is very wide.

Figures 4.1-9 show the electron density distribution in the second transistor. As was expected, the surface concentration is depressed because of the channel implantation. One can now observe a carrier channel at a depth of about 300 nanometers, which is also the depth of the pn-junction. This carrier channel was caused by way of the "punch through" effect, which was already observed from the potential distribution and which will be made even more clear by the current density distribution.

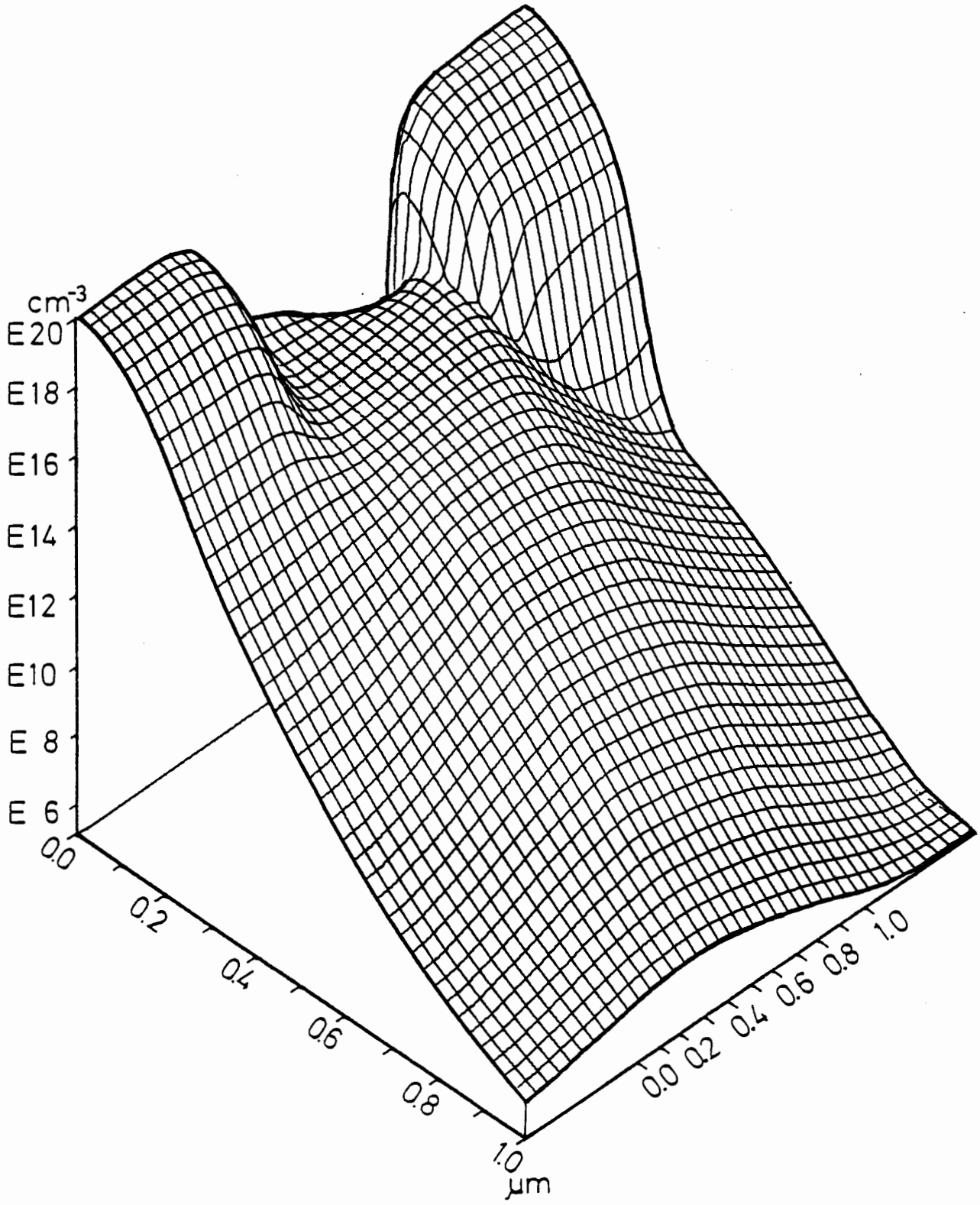


Figure 4.1-8a: The electron distribution in the first transistor.

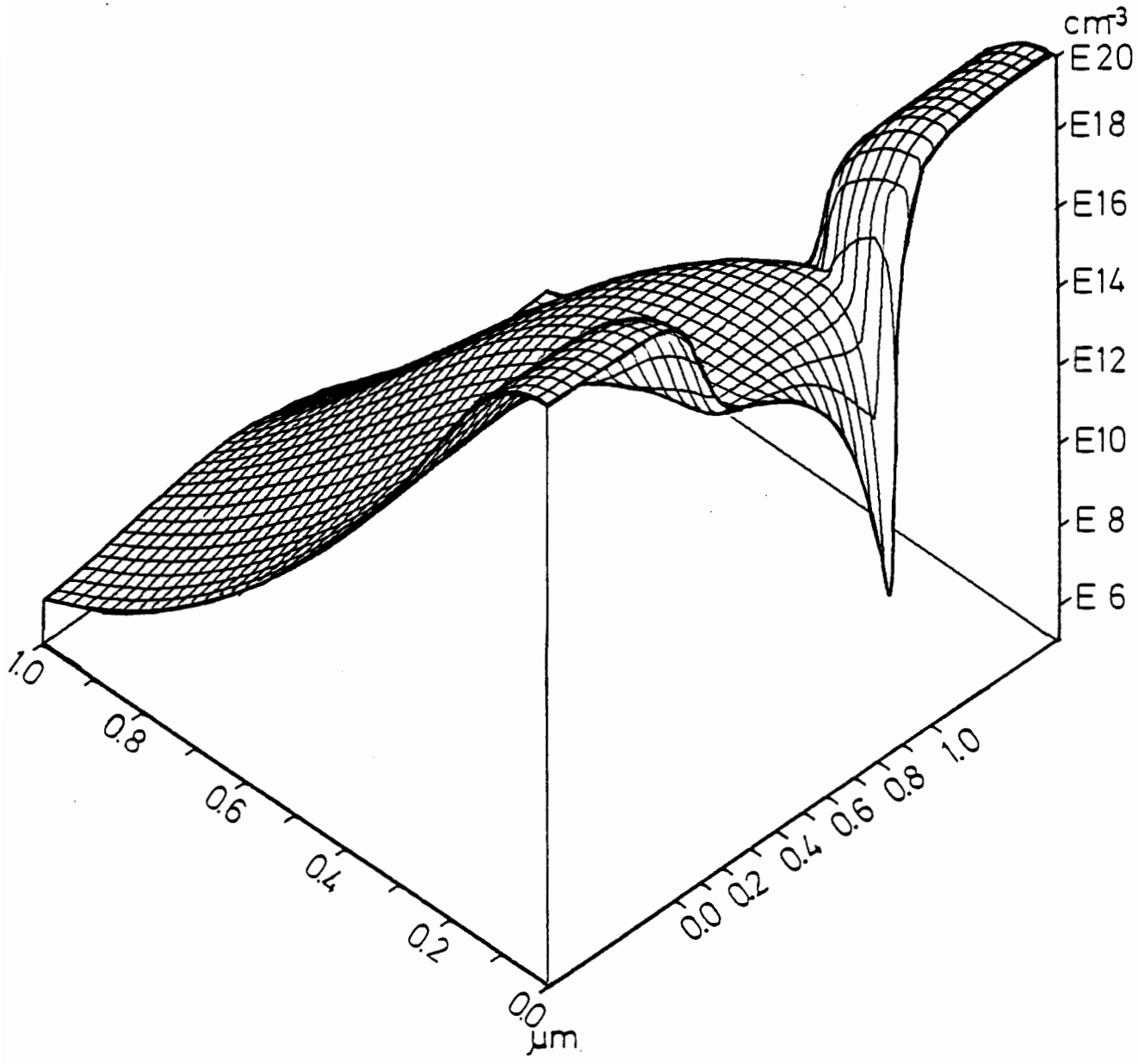


Figure 4.1-8b: The electron distribution in the first transistor.

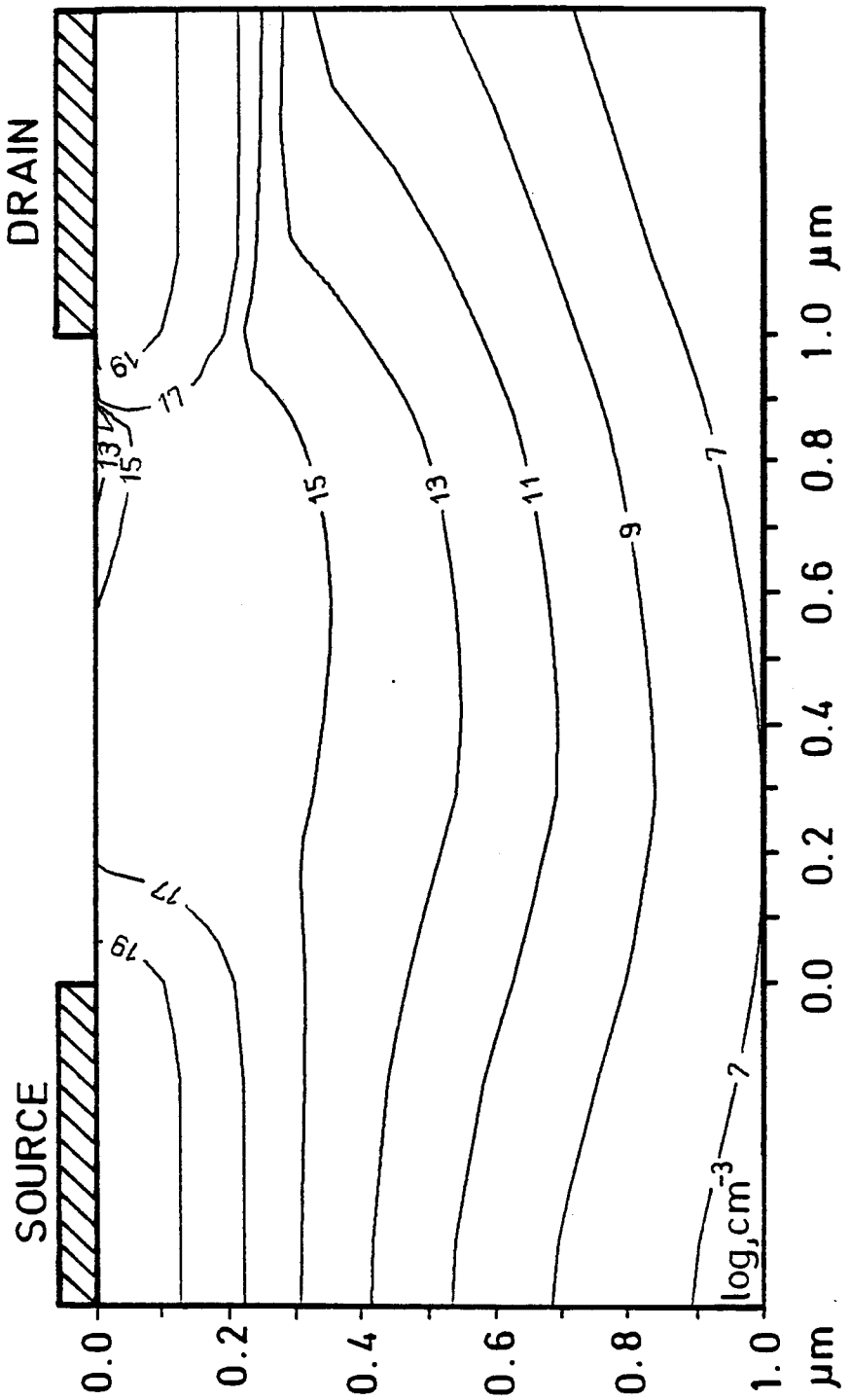


Figure 4.1-8c: The electron distribution in the first transistor.

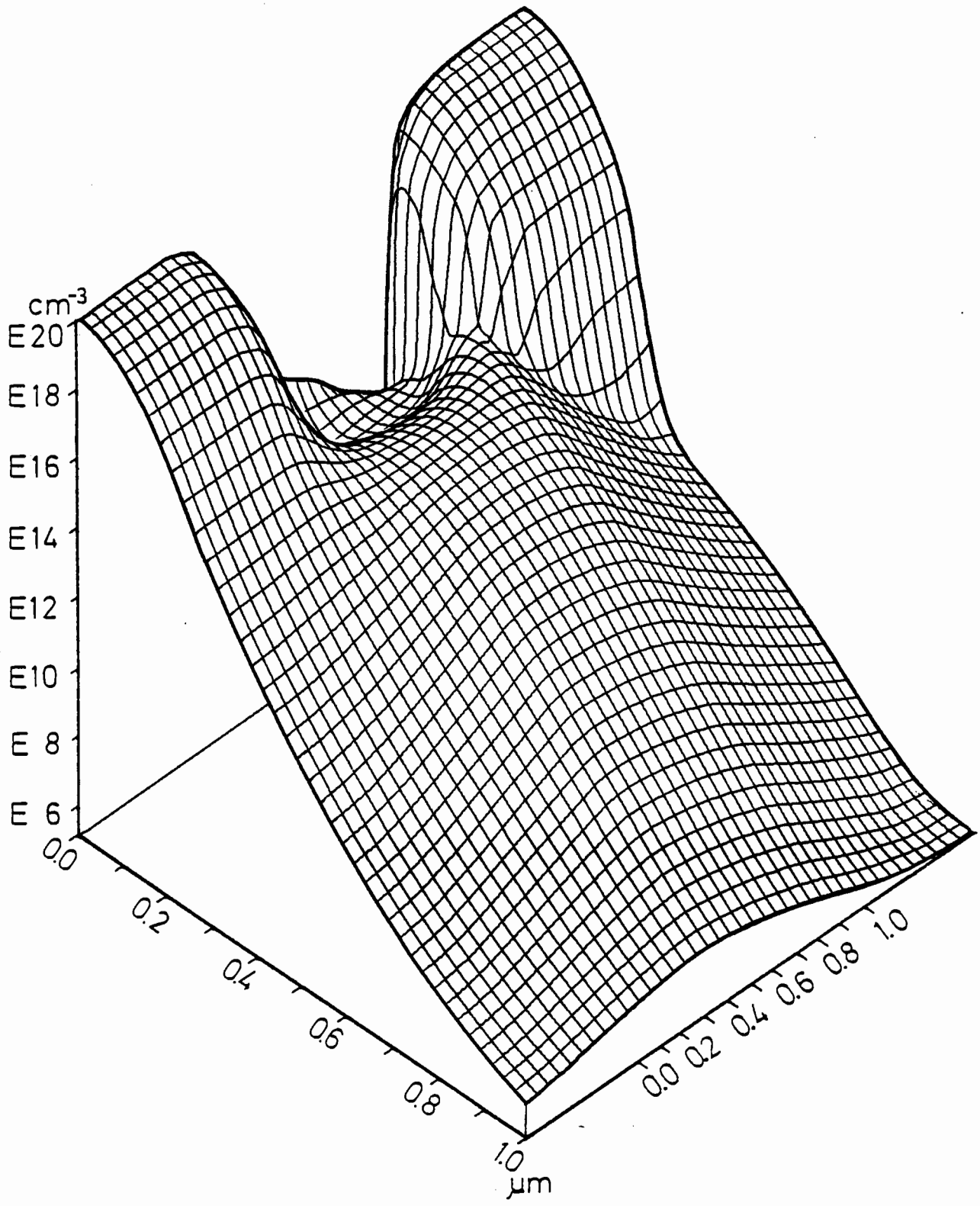


Figure 4.1-9a: The electron distribution in the second transistor.

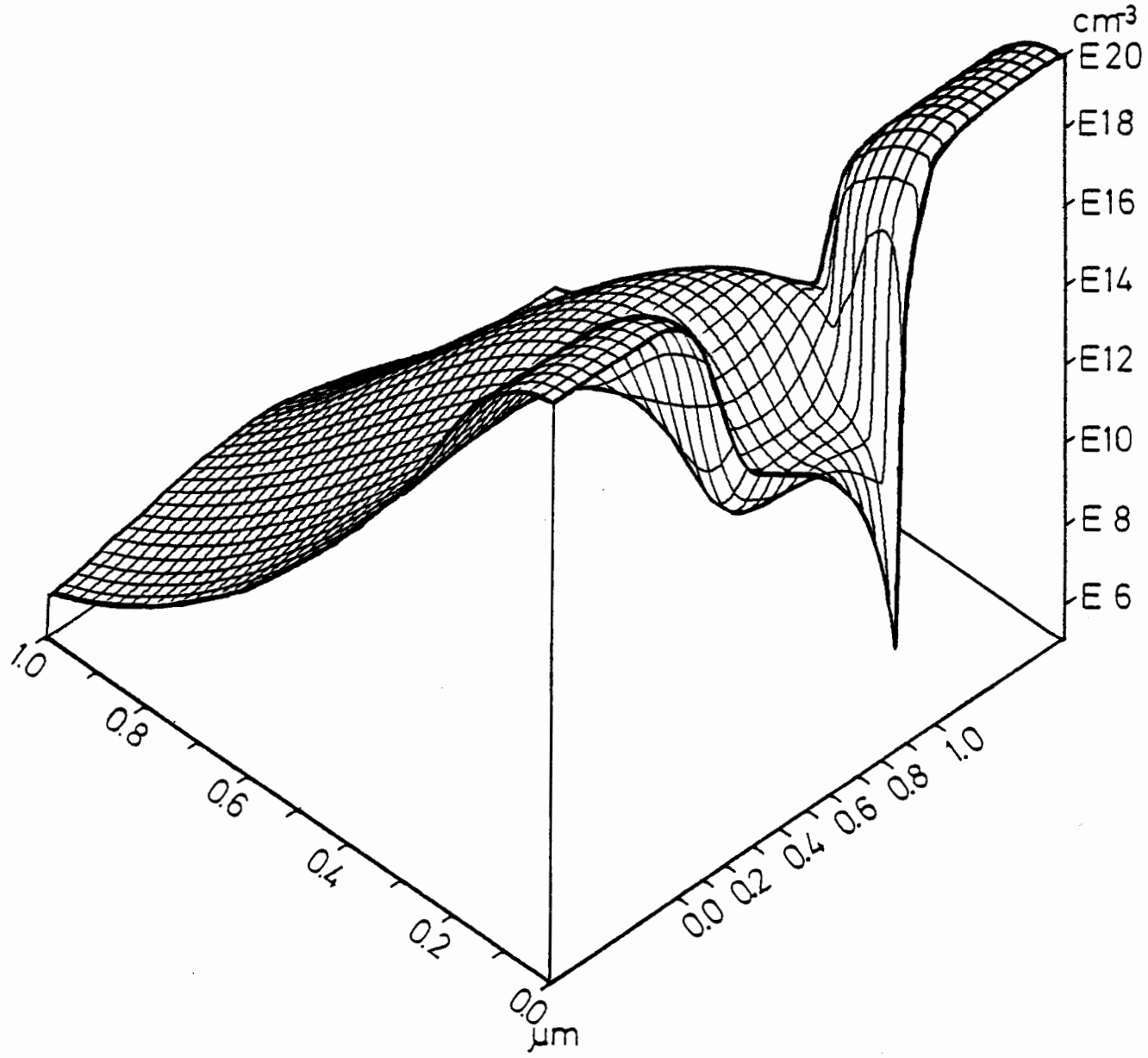


Figure 4.1-9b: The electron distribution in the second transistor.

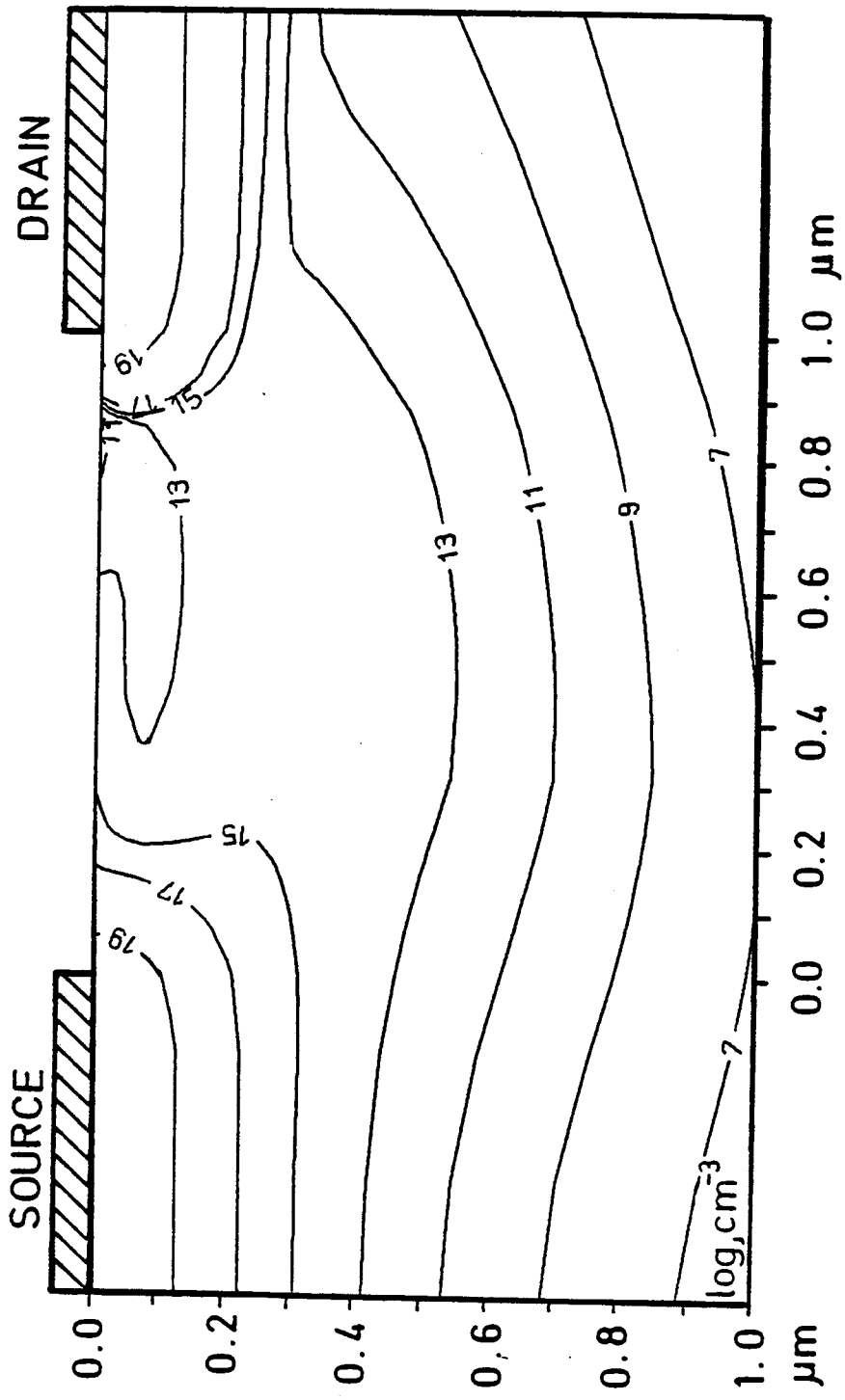


Figure 4.1-9c: The electron distribution in the second transistor.

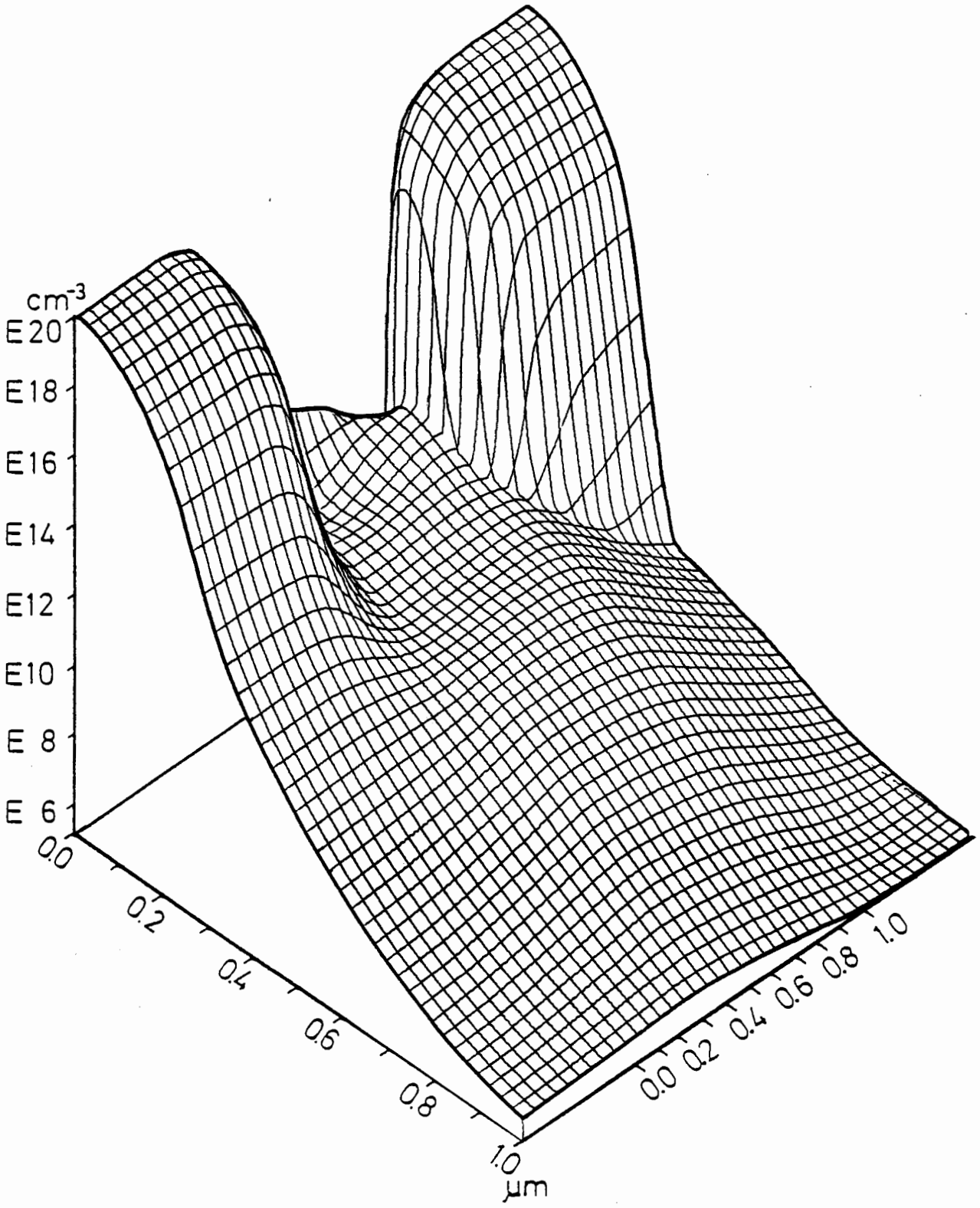


Figure 4.1-10a: The electron distribution in the third transistor.

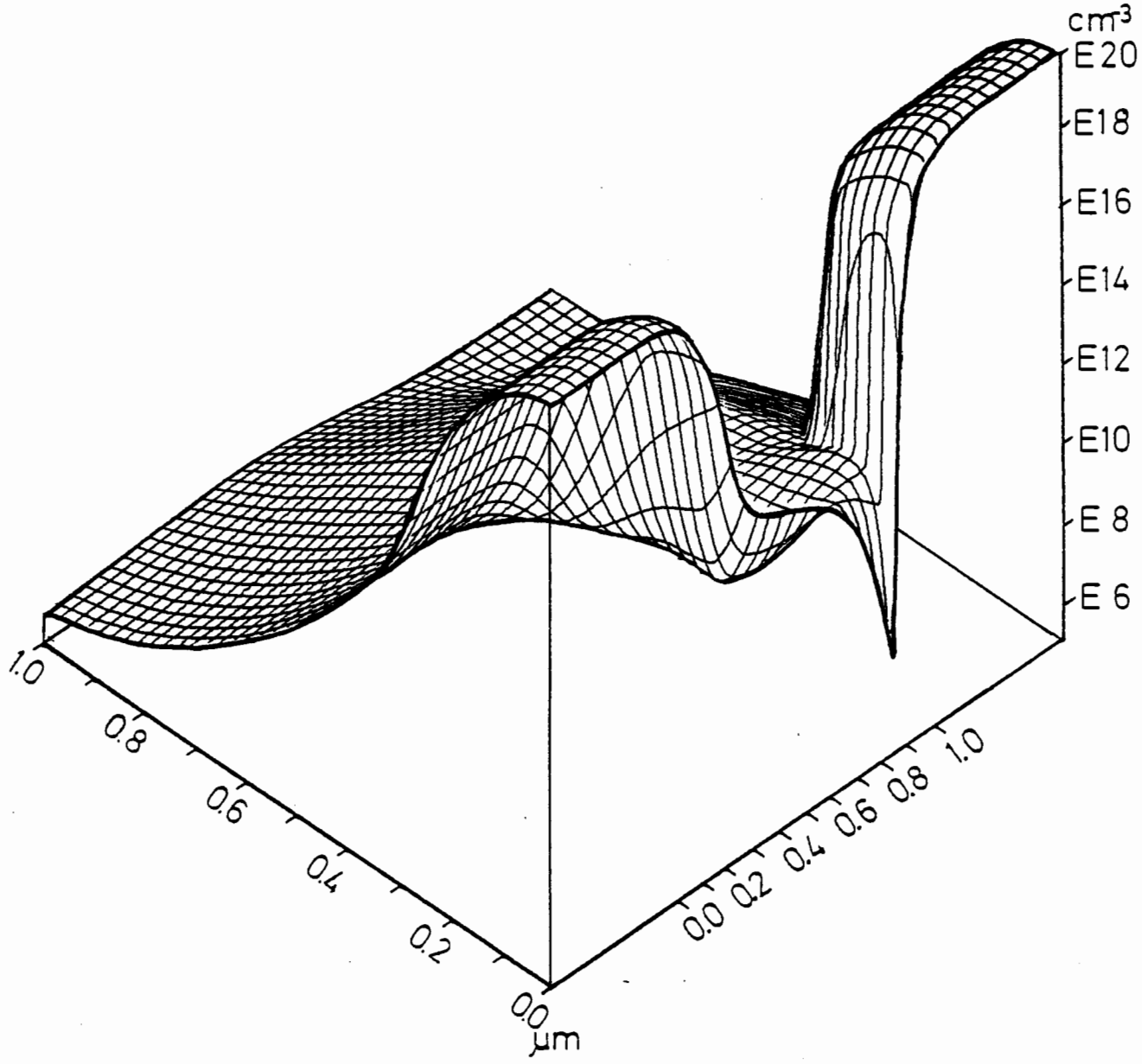


Figure 4.1-10b: The electron distribution in the third transistor.

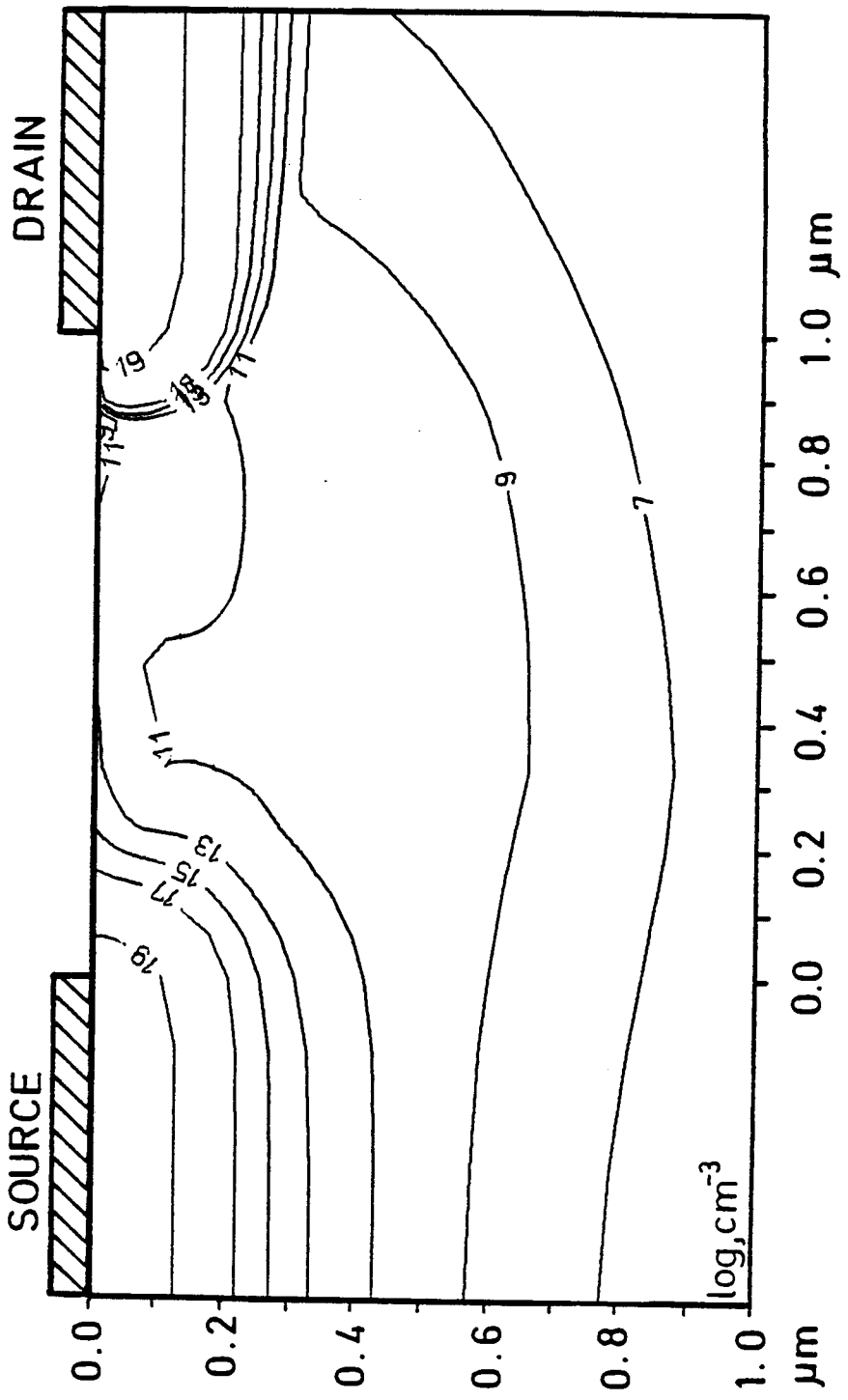


Figure 4.1-10c: The electron distribution in the third transistor.

Figures 4.1-10 show the electron density distribution in the third transistor. The second deep channel implantation causes the carrier concentration to fall off monotonically from the surface inside the transistor, which distinctly establishes the suppression of the "punch through" effect. Examining the absolute value of the carrier concentration it is immediately obvious that the operating point which was selected by a predetermined convention is not in the region of strong inversion, but lies in the region of deep depletion. It is also worth noting that the carrier concentration is beginning to separate relatively rapidly near the middle of the transistor. On the contrary the characteristic "pinch off" zone is very small and lies very near the drain region.

Figure 4.1-11a shows the distribution of the lateral current density component in the first transistor in a 3-D representation. In figure 4.1-11b the same quantity is shown as viewed from the surface in order to better view the channel region. At the source end of the channel the transverse field component forces the current to flow near the surface. However, in the middle of the channel the current flux is spread out under the influence of the drain voltage, which is a typical short channel effect. The channel itself is relatively wide. The reason for this is found to be a superposition of the inversion current and the "punch through" current. The maximum values of the lateral current density component lie, surprisingly, directly under the contacts. This becomes obvious when one looks at the considerations for current continuity. The current flows through the contact almost completely in the form of a transversal component. There is only a small lateral voltage drop under the contacts which can cause only a small lateral current component. In the source/drain regions the lateral current density component must increase immediately in order that a lateral current flow will result through the transistor. Because of the source free nature of the current, the integration (in a direction normal to the surface) of the lateral current density component must result in a flux which is constant anywhere in the channel region (conservation of flux). This naturally results in high current densities in narrow channel regions and lower current densities in wide channel regions.

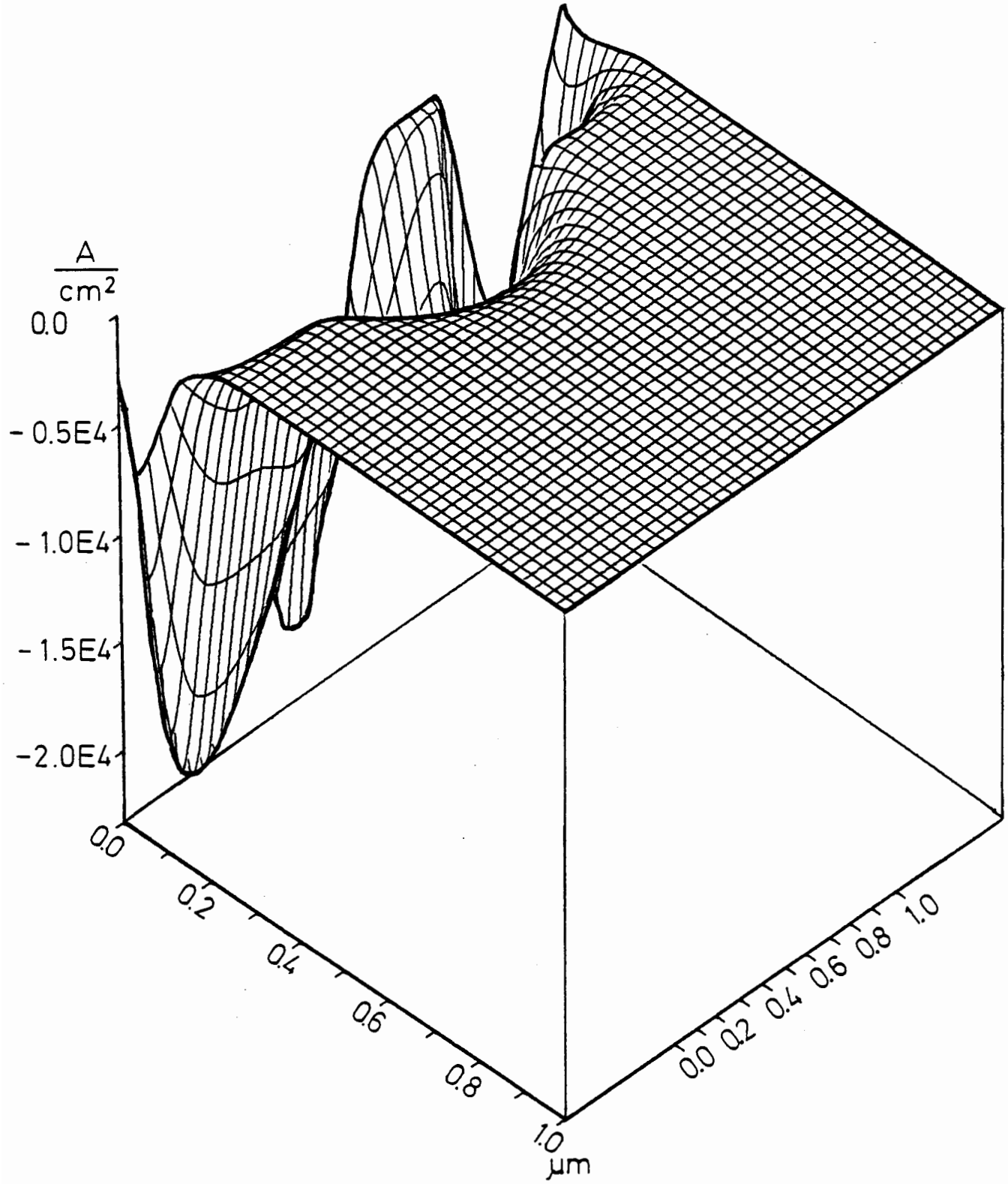


Figure 4.1-11a: The current density distribution in the first transistor.

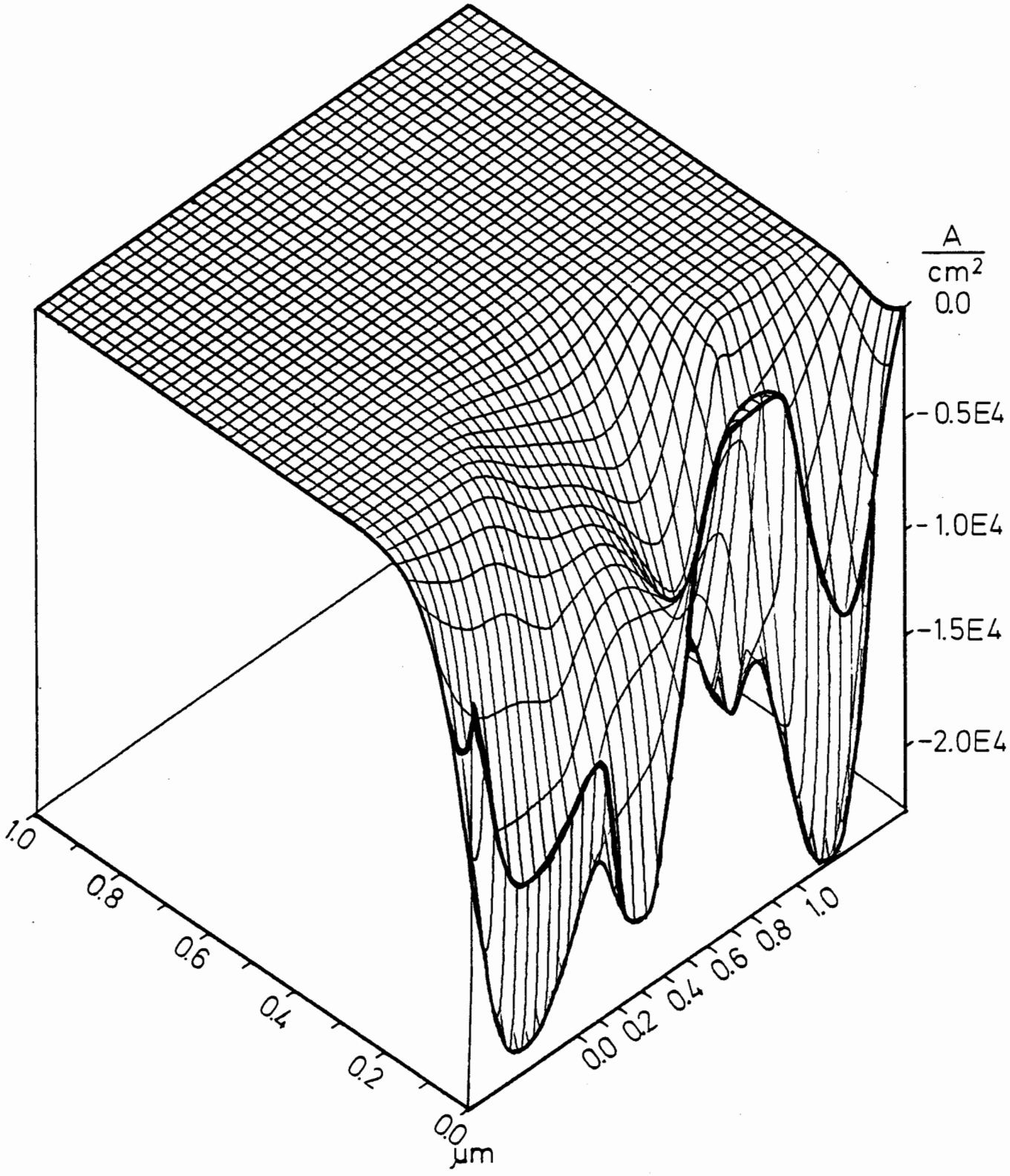


Figure 4.1-11b: The current density distribution in the first transistor.

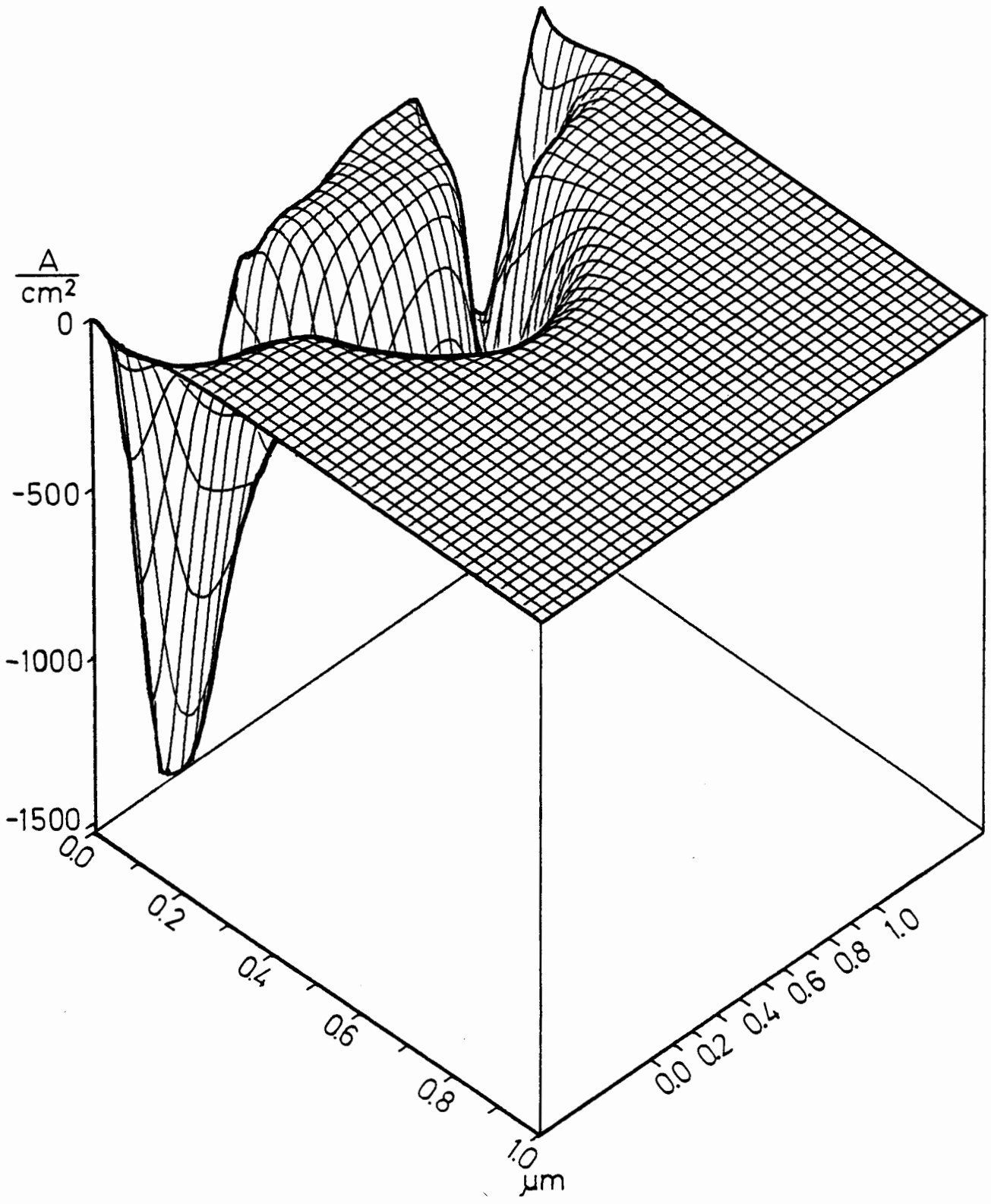


Figure 4.1-12a: The current density distribution in the second transistor.

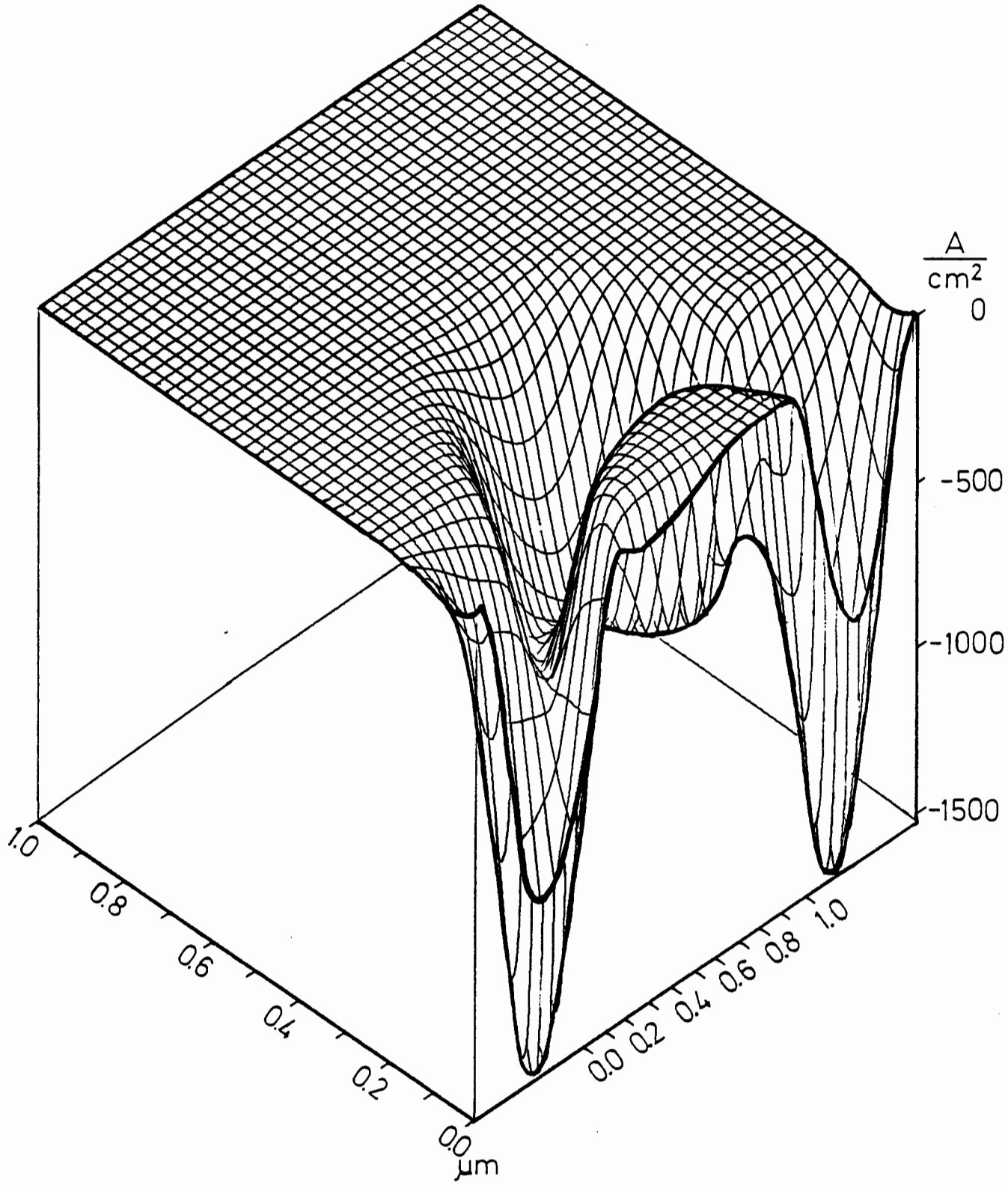


Figure 4.1-12b: The current density distribution in the second transistor.

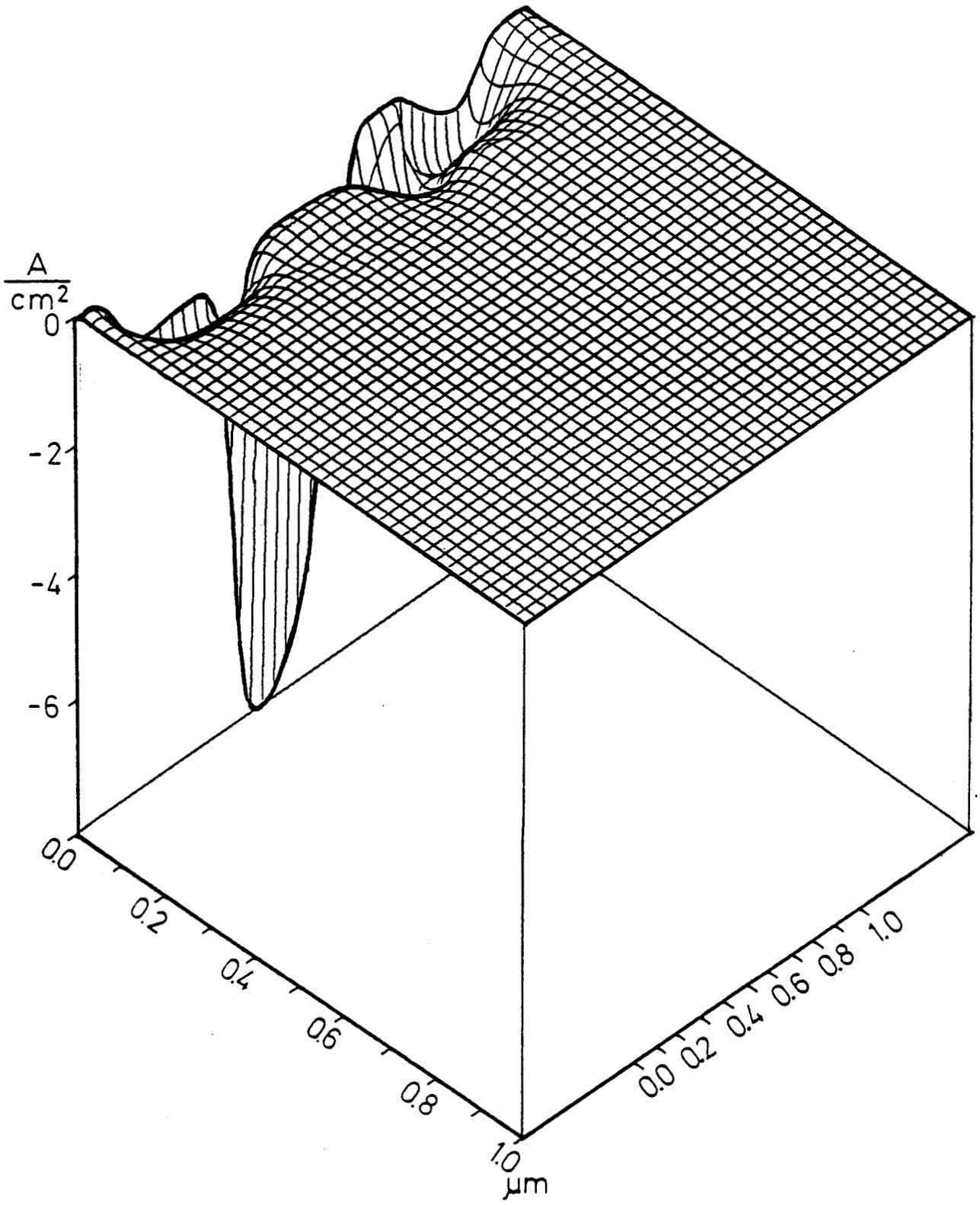


Figure 4.1-13a: The current density distribution in the third transistor.

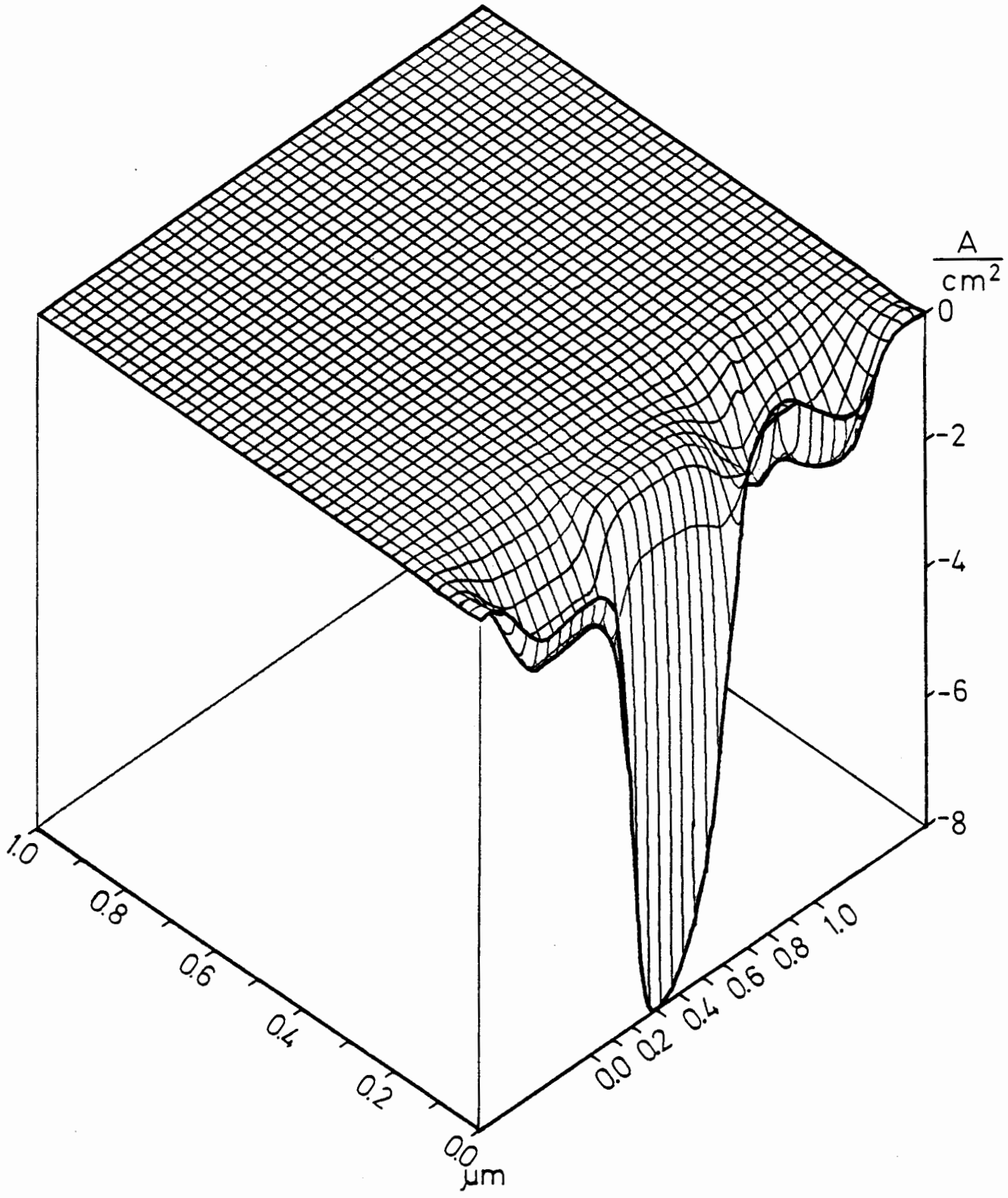


Figure 4.1-13b: The current density distribution in the third transistor.

The distribution of the lateral current density component in the second transistor is shown in figures 4.1-12. One observes a massively marked "punch-through" effect. The current flux occurs in a wide channel in the substrate /32/. Practically no portion of a pure surface current is present here. The peak current density at this operating point is, in comparison to the first transistor, an order of magnitude smaller.

Figures 4.1-13 show the distribution of the lateral current density component in the third transistor. The "punch-through" channel of the last figure has totally disappeared. The total current flux occurs completely at the surface. The peak current density -which appears here in the middle of the channel and directly at the surface- is about a factor of 200 smaller in comparison to the second transistor. Current density distributions with these qualitative characteristics are typical for ordinary functional devices in the subthreshold region and can be used as very good evaluation criteria.

Figure 4.1-14 shows the mobility distribution in the first transistor in the form of isocontours. In the highly doped source/drain regions the mobility is very small because of impurity scattering. Under the source region the mobility immediately increases to its bulk value. This is naturally not the case under the drain region, because here the strong fields in the reverse biased drain/substrate diode decrease the mobility by way of the saturated drift velocity. The local maximum in the mobility under the drain region is worth noting; this results because there is only a small amount of impurity scattering and also the field strength is not high enough in order to decisively reduce the mobility. This effect will also be seen in all of the following mobility distributions. The mobility in the channel falls off monotonically along the surface. Only in the region of the channel near the source does it appear constant. In this short region the field strength component parallel to the current direction is very small and this results in no reduction in the mobility. This field strength component increases nearer the drain, which results in the above mentioned mobility reduction. In the normal direction the mobility parallel to the surface is reduced by surface scattering. An exact explanation of the modeling of the different scattering mechanisms and the mathematical formulas used are found in section 2.2.3.

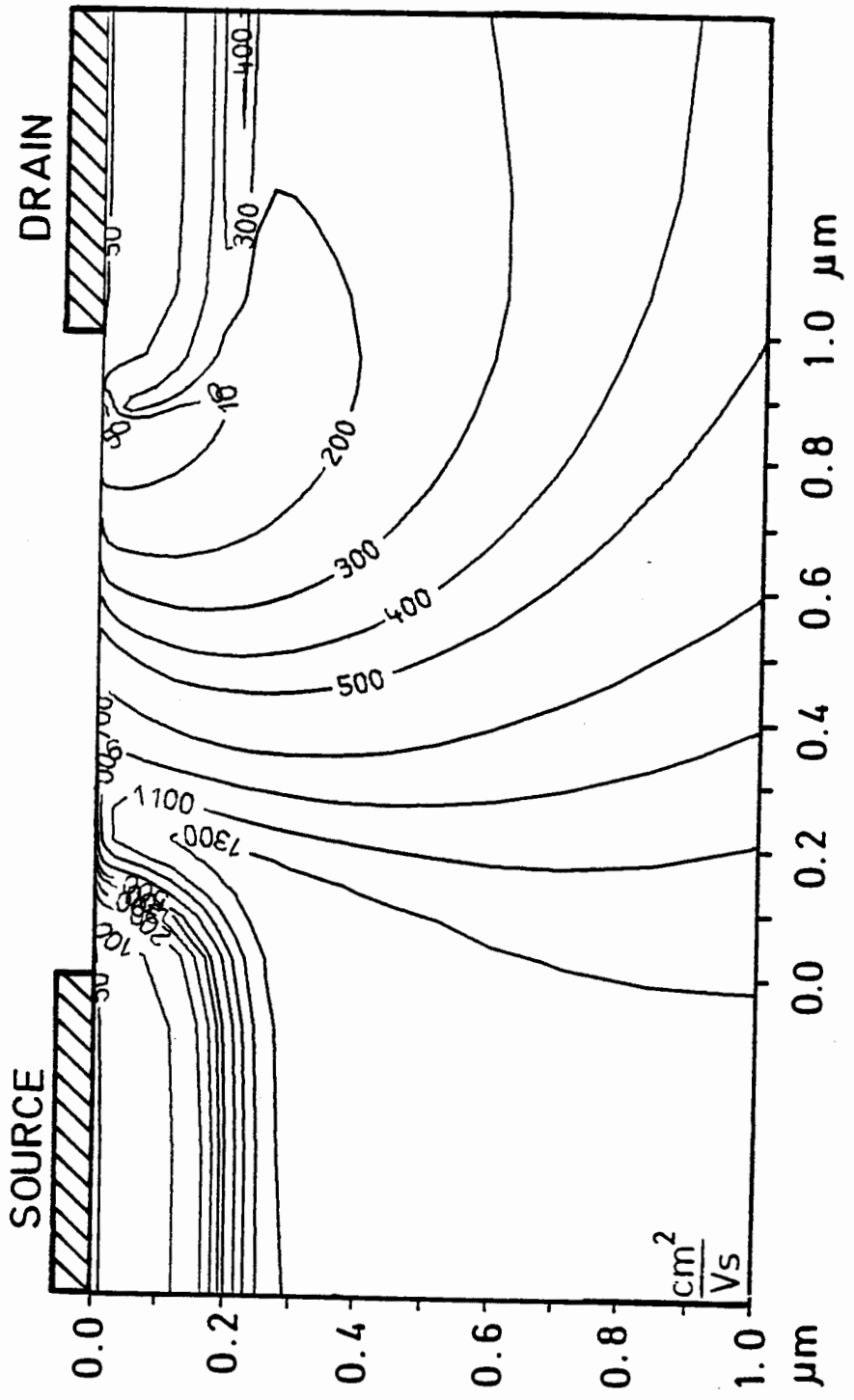


Figure 4.1-14: The mobility distribution in the first transistor.

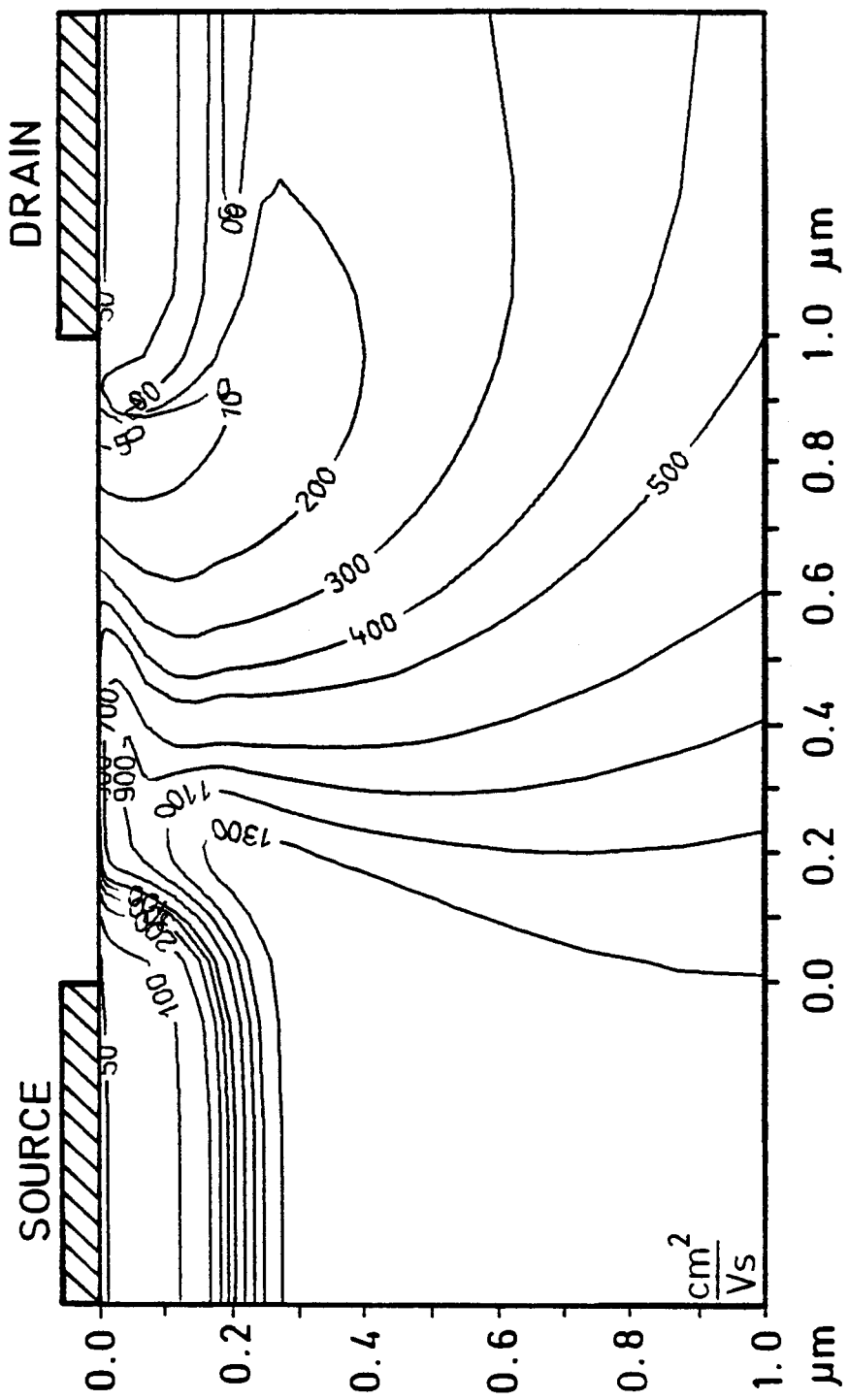


Figure 4.1-15: The mobility distribution in the second transistor.

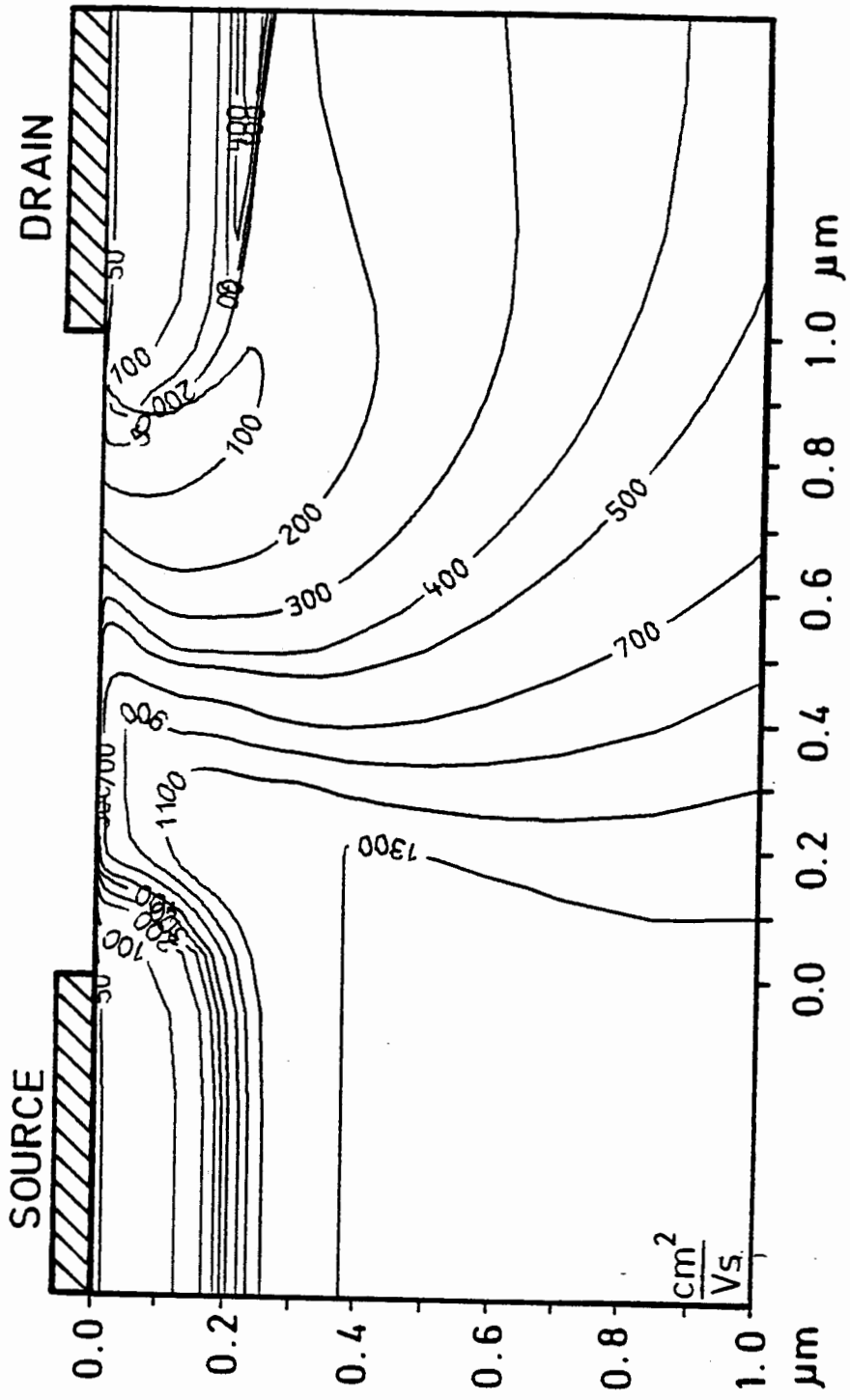


Figure 4.1-16: The mobility distribution in the third transistor.

Figure 4.1-15 shows the isocontours of the mobility in the second transistor. The qualitative appearance of this figure is, as was expected, relatively similar to figure 4.1-13. One distinctly observes the appropriate mobility reduction at the surface due to scattering by the impurities of the channel implantation. It is worth noting that the region along the surface in which the mobility is negligibly reduced is somewhat longer here. This results because of the suppression of the short-channel effects by way of the channel implantation.

Figure 4.1-16 shows the mobility distribution in the third transistor. In the channel region there is essentially no change in comparison with the second transistor. Only at the depth of the p-n junction can there be noticed a small mobility reduction due to the impurities of the second implantation. This effect has no influence upon the device's behavior.

Figure 4.1-17 shows the subthreshold characteristics for the three transistors presented here for two different drain voltages. The solid curves correspond to 100 millivolts and the dashed curves correspond to 3 volts. The slope is identical for all three transistors at 100 millivolts. At 3 volts the slopes are substantially reduced for the cases of the first and second transistors by way of the "punch through" current. In the third transistor the displacement of the characteristic curves at different drain voltages due to short channel effects is minimal.

Fundamental statements about the behavior of a transistor can be derived from the subthreshold characteristics of short channel transistors as can be seen from the above example. One can apply them directly toward the definition of the threshold voltage, and, what is of greater importance, their slope immediately gives an indication of the usefulness of the transistor in a circuit /38/. If the slope is not large enough such that the transistor cannot be reasonably turned off, the resulting leakage current in the circuit will be unacceptable.

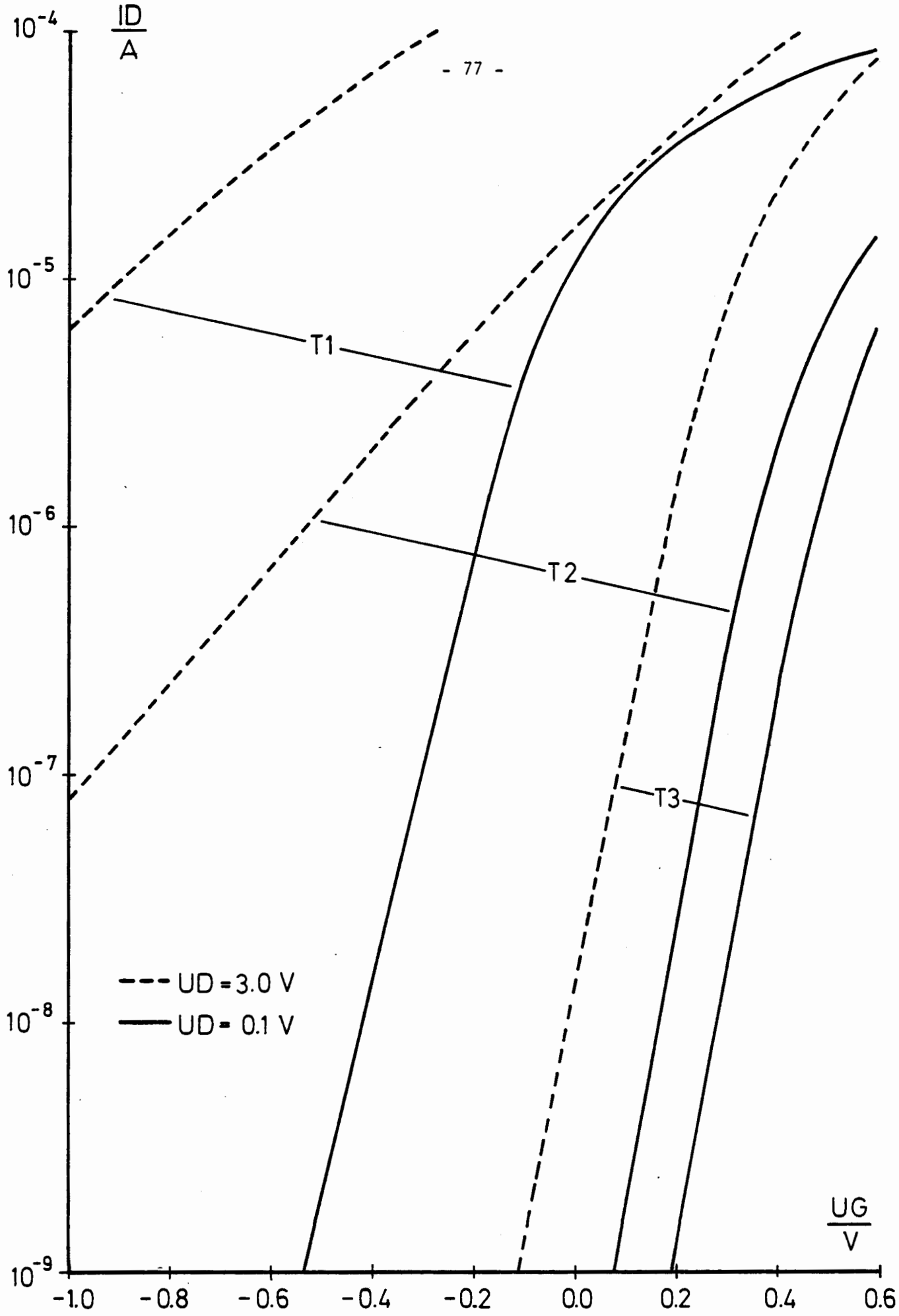


Figure 4.1-17: Subthreshold characteristics.

4.2 The Simulation of an Inverter

The simple inverter is one of the basic building blocks of digital integrated circuits. Such a simple inverter consists of two types of devices; a drive transistor and a load element. In the case of a low level on the gate the drive transistor should draw no current, thereby no voltage drop will occur across the load element and consequently the drain contact, which represents the output of the inverter, will be at a high level. The application of a high level on the gate should turn the drive transistor on completely and consequently a low level should appear at the output due to the voltage drop across the load element. Thereby one obtains the desired electrical behavior of the inverter by placing certain restrictions on the physical behaviors of the drive transistors and the load elements. The threshold voltage of the drive transistor, for example, must be greater than the low level. Many possibilities are available for the load element. It can be realized as an ordinary ohmic resistor or as a load transistor. The current which a modern miniaturized drive transistor can draw is relatively small. The power dissipation should remain small. Therefore it would be difficult to implement the load element as an ohmic resistor in view of the high resistance which is required. Such a resistor which would be represented in practice by a diffused region would be very large and the gain in packing density which resulted from the miniaturization of the active element, the drive transistor, would be relatively small.

Presently, because of the above mentioned reasons, active elements are used almost completely as loads. Such a load transistor can be implemented in one of two possible ways: in the saturation region and/or unsaturated in the ohmic region. The condition for operation in the ohmic region is, that the applied gate/source voltage (UGS) must always be greater than the sum of the threshold voltage (UT) and the drain/source voltage (UDS).

Figure 4.2-1 shows a typical inverter circuit with an active load. T1 is the drive transistor, T2 is the load transistor. One selects a transistor with a negative threshold voltage for T2, such that it always operates in the ohmic region, which on the one hand results in a larger high/low-level swing and on the other hand contributes to increasing the speed of the circuit/63/.

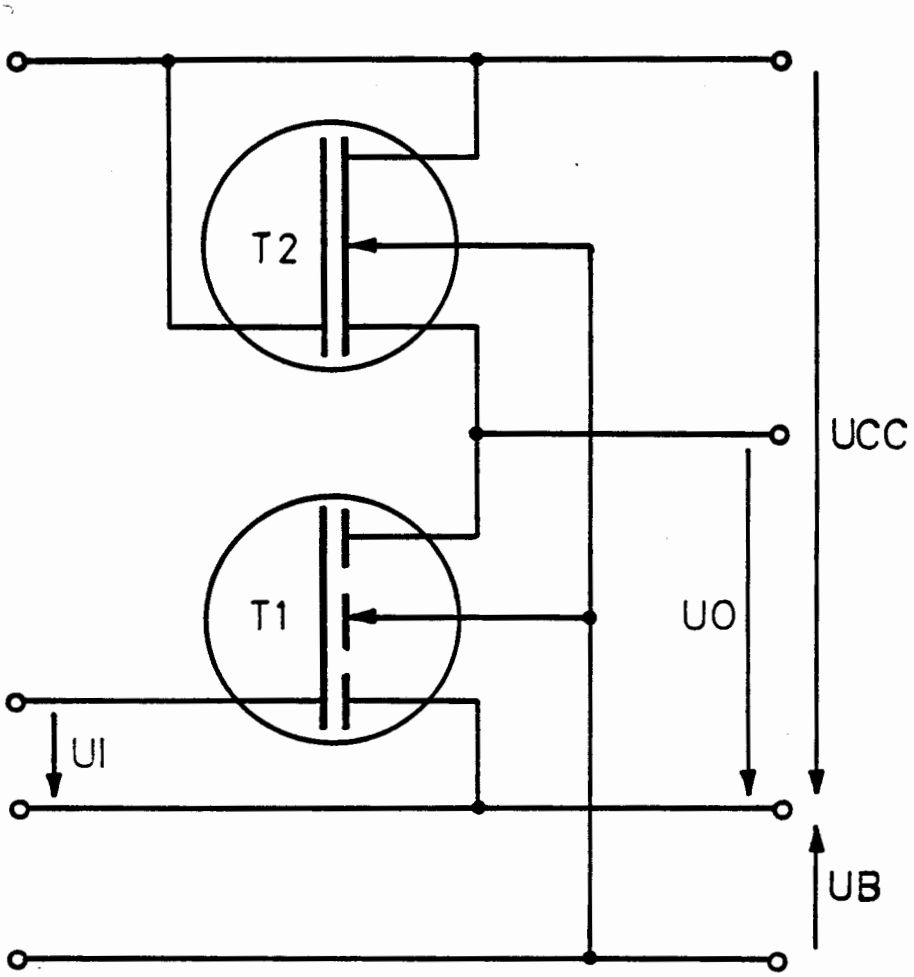


Figure 4.2-1: A simple inverter.

The required negative threshold voltage can be obtained through technological means. In the following only one of the transistors which fulfilled the requirements for a load, which together with the third transistor of subsection 4.1 can be used in an inverter circuit, will be investigated.

4.2.1 The Load Transistor

One uses a donor element for the channel implantation, such that a slight connection is made between the source and drain, without which one must create an inversion channel by way of the gate voltage; in practice the channel will be implanted. By proper design for the channel implantation, the controllability of the current flow in the channel by way of the gate is not lost, instead only the threshold voltage is shifted by a negative amount.

Figure 4.2-2 shows the doping profile of the load transistor. One distinctly observes the steep donor implantion which was carried out for the above reasons, in order to obtain a built-in channel. The implantation was carried out with antimony, a dose of 10^{12} cm^{-2} and an energy of 180 keV. All other technological and geometrical data are identical to the input for the third transistor of subsection 4.1. The steep boron implantation of the above mentioned transistor has naturally fallen off.

In order to obtain a better feeling for and understanding of the behavior of a "depletion mode" transistor, the following is a short discussion of the distribution of the relevant physical quantities for the same operating point as was used in subsection 4.1. A drain voltage of three volts will be used with a gate and substrate voltage of zero volts each.

Figure 4.2-3 shows the isocontours of the electrostatic potential. There is no barrier between the source and channel regions. A source/channel diode is naturally non-existent and one can, in the best case, distinguish the change of the built-in potential as an n^+n transition. Figures 4.2-4 show the electron density distribution in the load transistor. The presentation is analogous to that of subsection 4.1. One distinctly observes a carrier channel, which extends over the entire channel region and the maximum concentration is at a depth of about 100 nanometers. Naturally, due to the

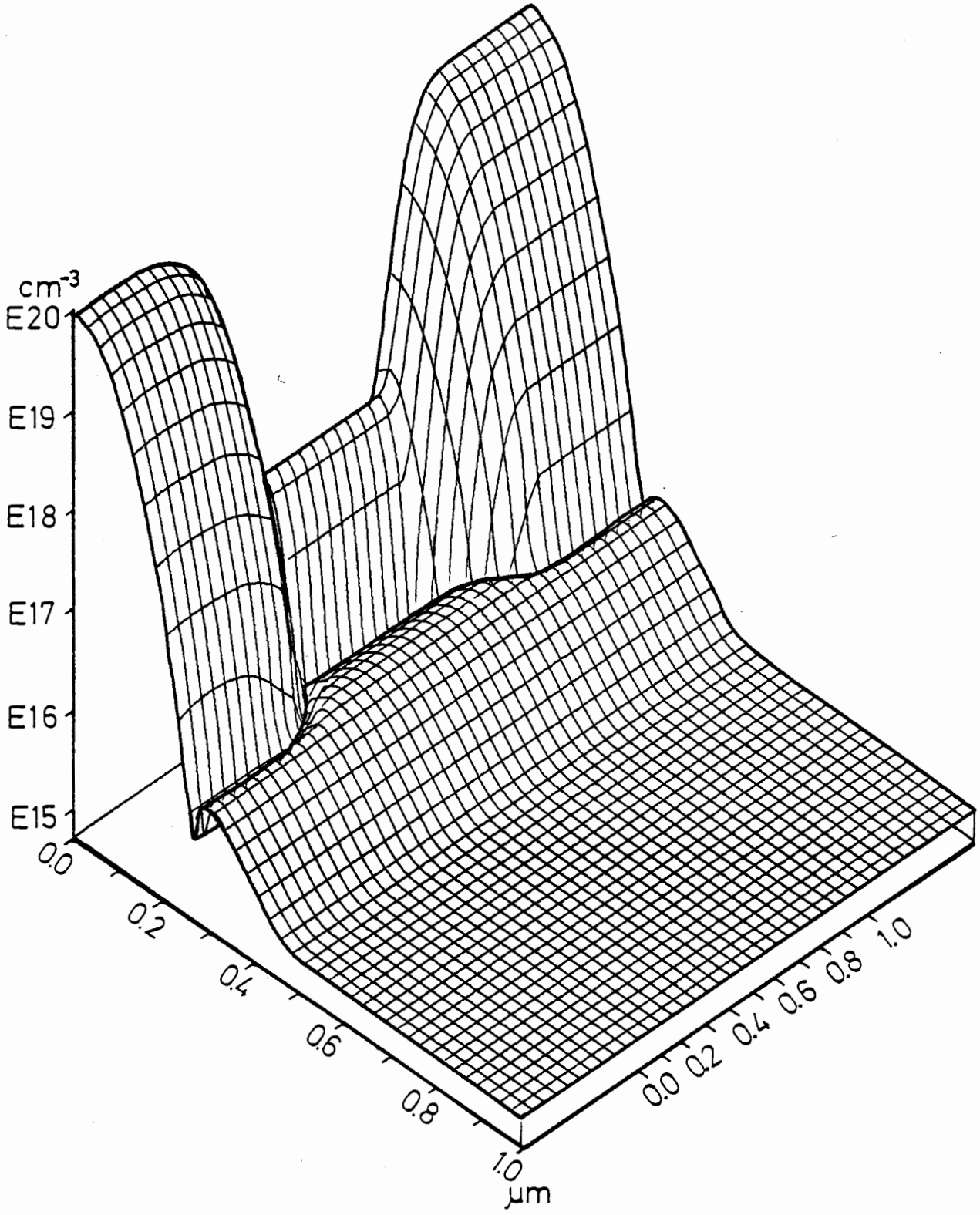


Figure 4.2-2: The doping profile of the load transistor.

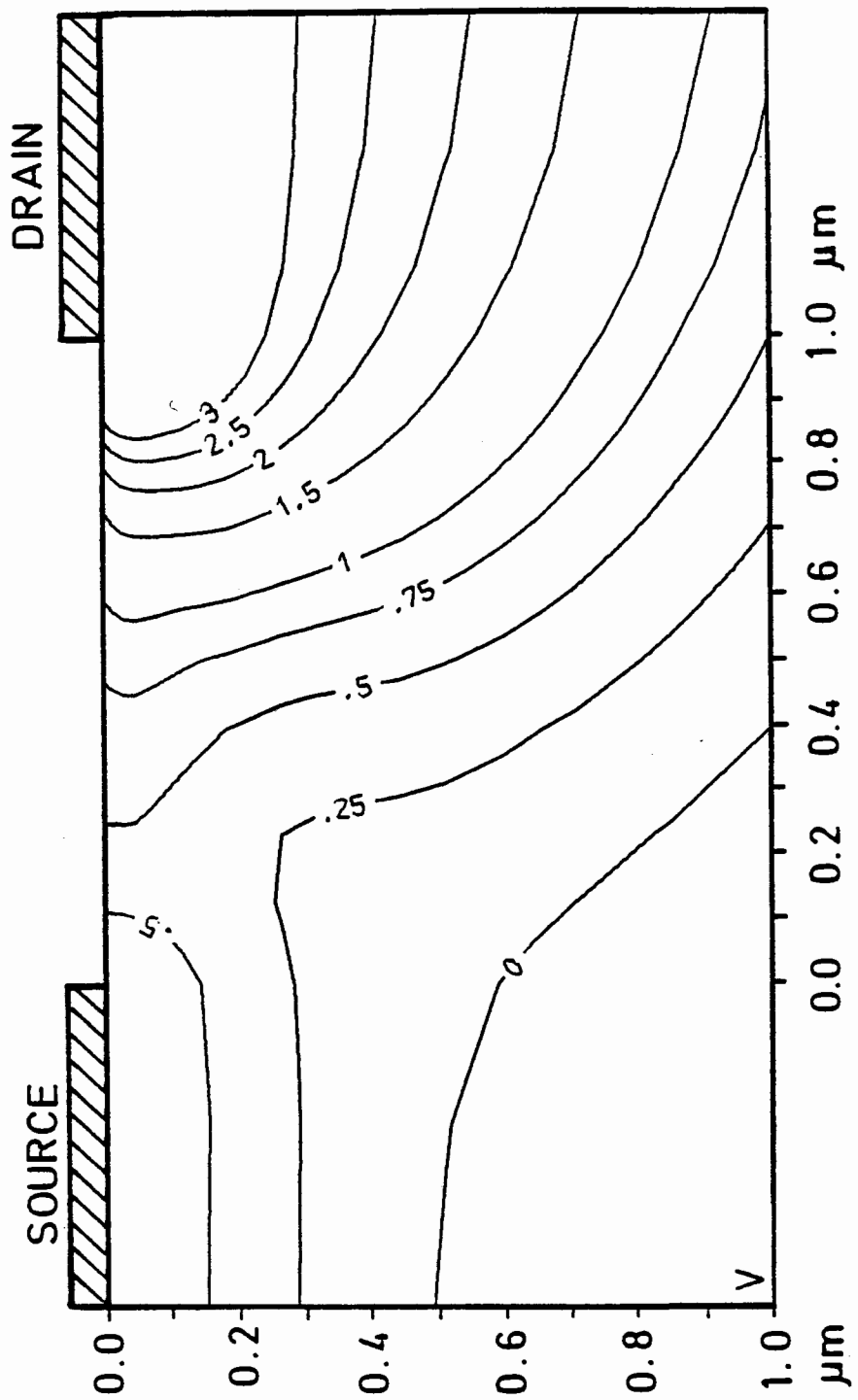


Figure 4.2-3: The potential distribution in a load transistor.

selected operating point, there exists a "pinch off" zone in this device.

Figures 4.2-5 show the distribution of the lateral current density component in the load transistor in the usual 3-D presentation. The channel is relatively wide and has its peak current density at a depth of about 100 nanometers, which was identically observed for the electron distribution. It is also worth noting that there is not the smallest indication of "punch through".

Figure 4.2-6 shows the mobility distribution in the load transistor in the form of isocontours. It is very similar to the distribution of the third transistor of subsection 4.1, the drive transistor in that subsection. It can be mentioned that the mobility was reduced by the higher dose in the steep implantation.

Figures 4.2-7 and 4.2-8 should definitely clarify a discussion of the internal behavior of the transistors. Figures 4.2-7 show the electron density distribution in the load transistor at a higher gate voltage of (UG=3V) and figures 4.2-8 show the same quantity for the drive transistor. Because of the higher gate voltage both transistors are operating in the region of strong inversion. In both transistors one can observe an enormous density gradient at the surface. The pure inversion channel of the drive transistor is thinner because of the antimony implantation defined channel of the load transistor. Because of this fact the same peak current density in the channel of the load transistor can furnish many times the total current of the drive transistor.

4.2.2 The Transfer Function

A nonsaturating inverter was analysed using the transistor discussed in the last section and the transistor which was discussed in the beginning of the chapter. The channel width of the drive transistor was set at 20 micrometers, and the channel width of the load transistor was set at 2.5 micrometers, in order to obtain a desired quotient for the control factor /59/. The substrate voltage was set at two volts, such that the threshold of the driver is high enough for a useful low level signal to noise ratio.

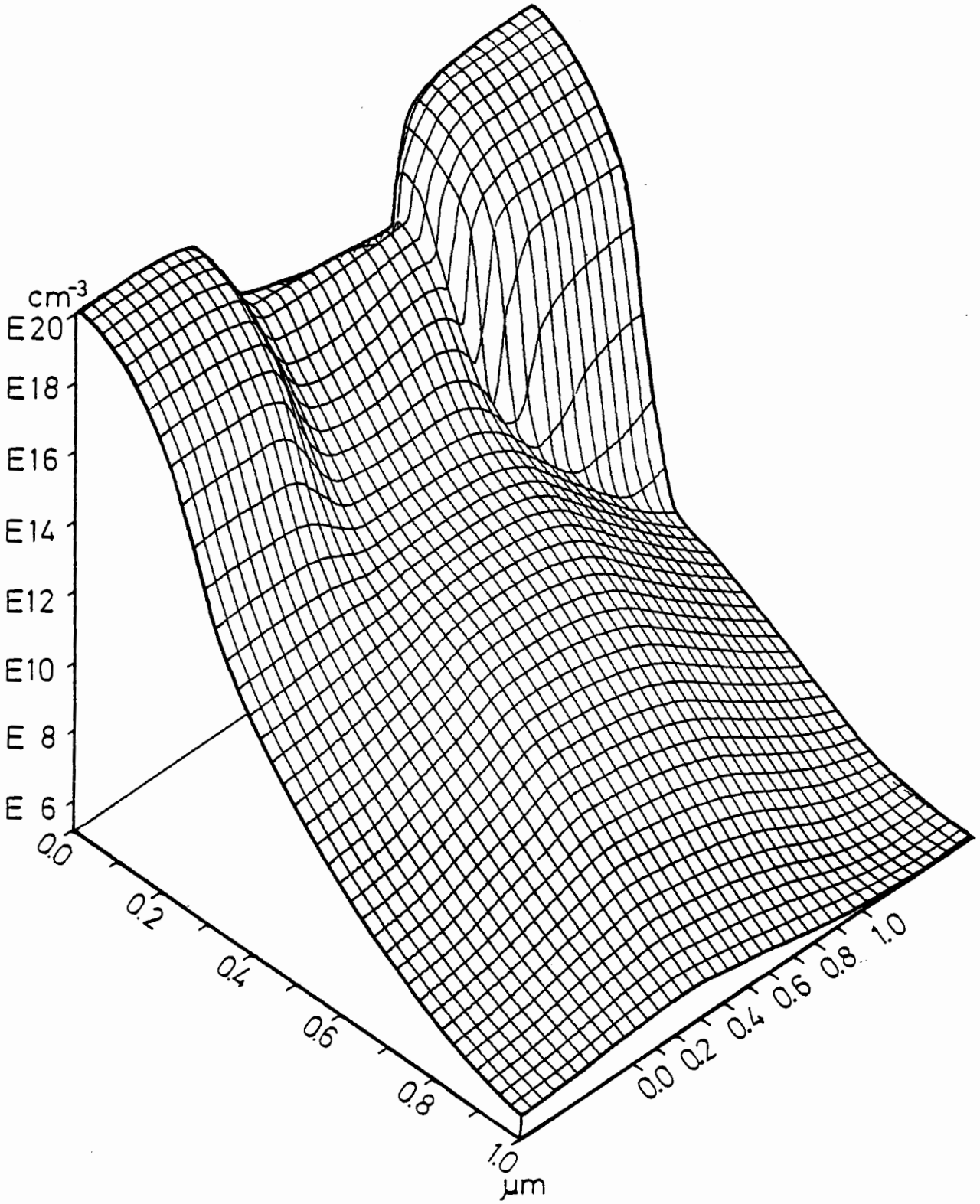


Figure 4.2-4a: The electron distribution in a load transistor.

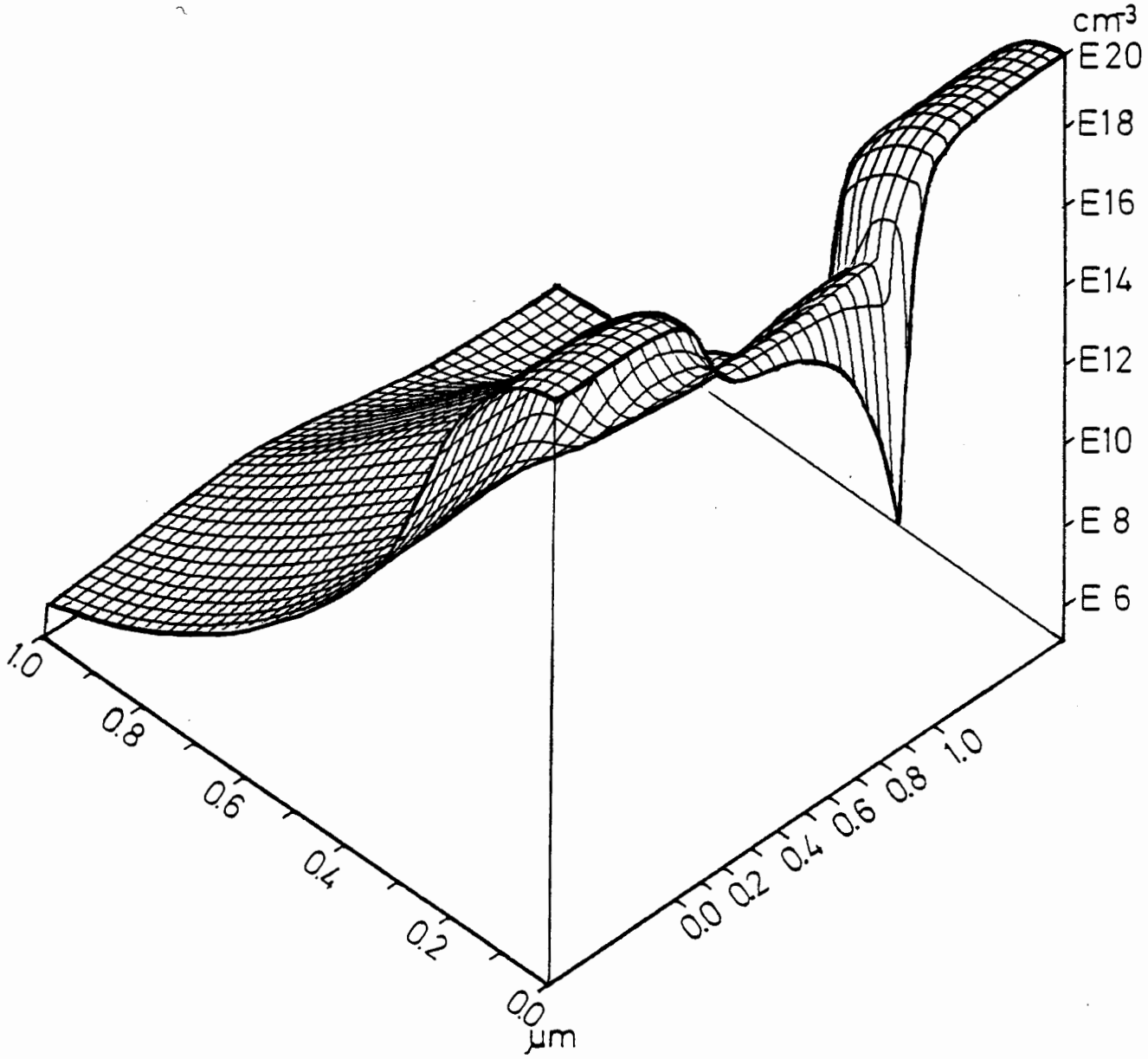


Figure 4.2-4b: The electron distribution in a load transistor.

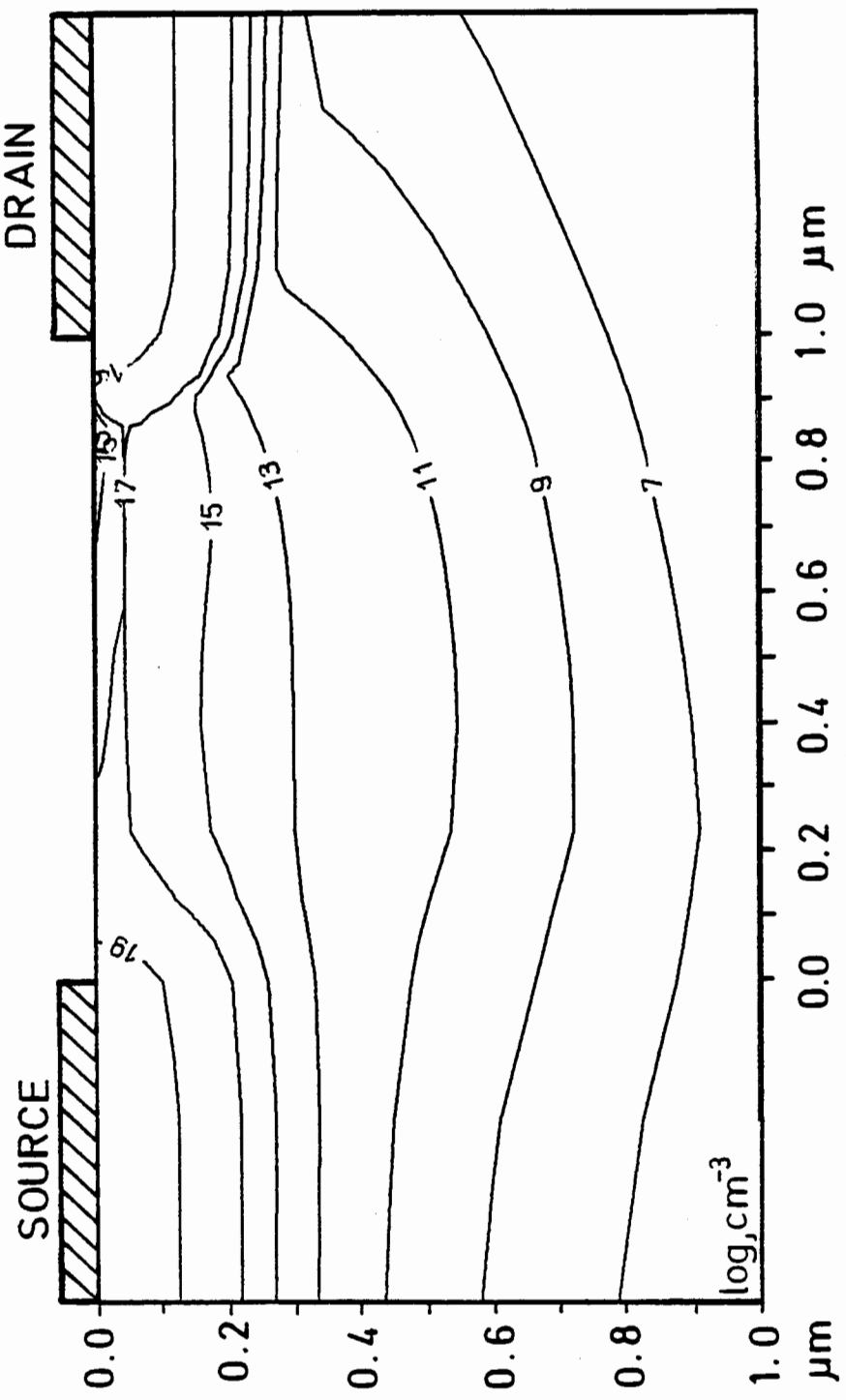


Figure 4.2-4c: The electron distribution in a load transistor.

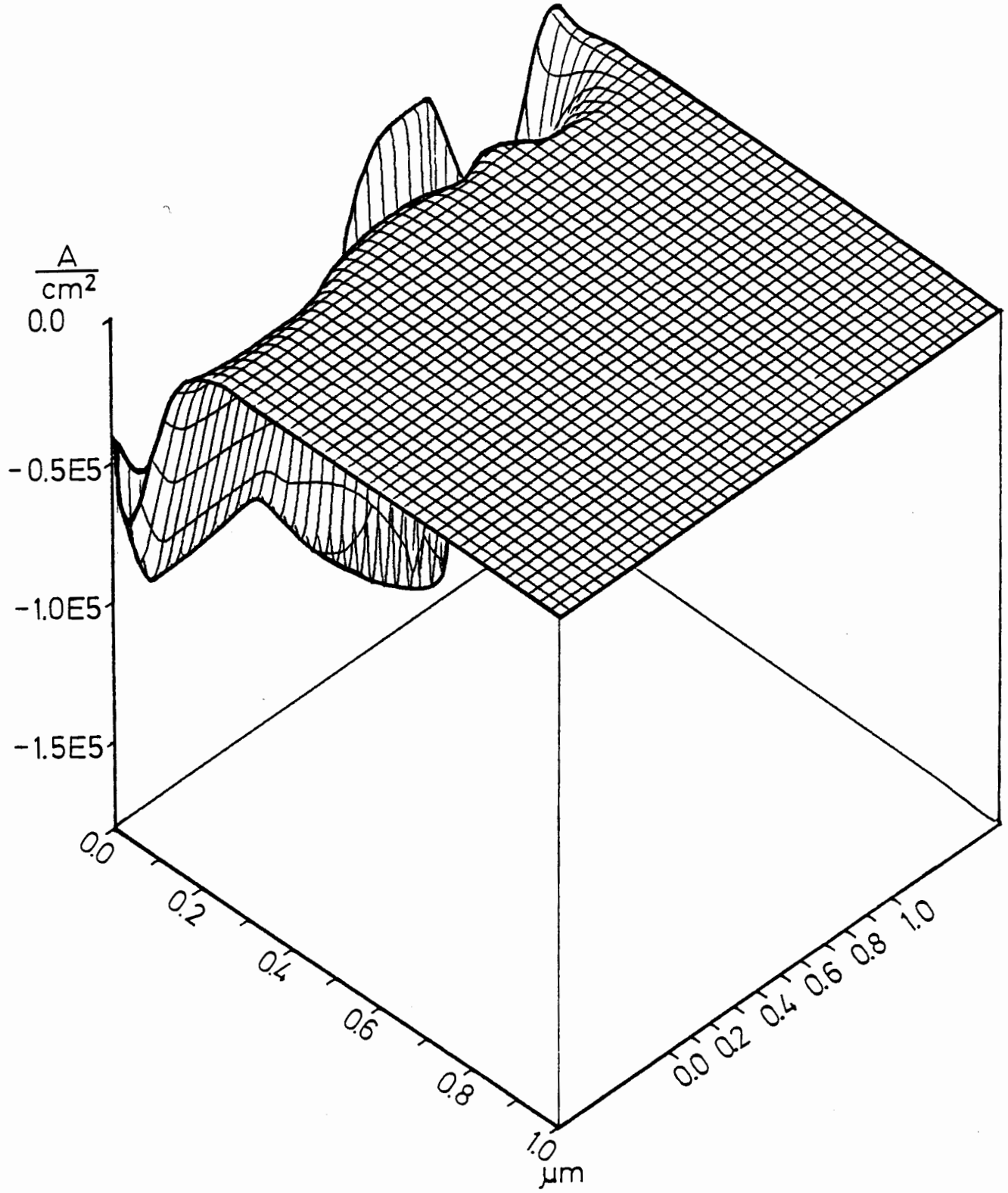


Figure 4.2-5a: The current density distribution in the load transistor.

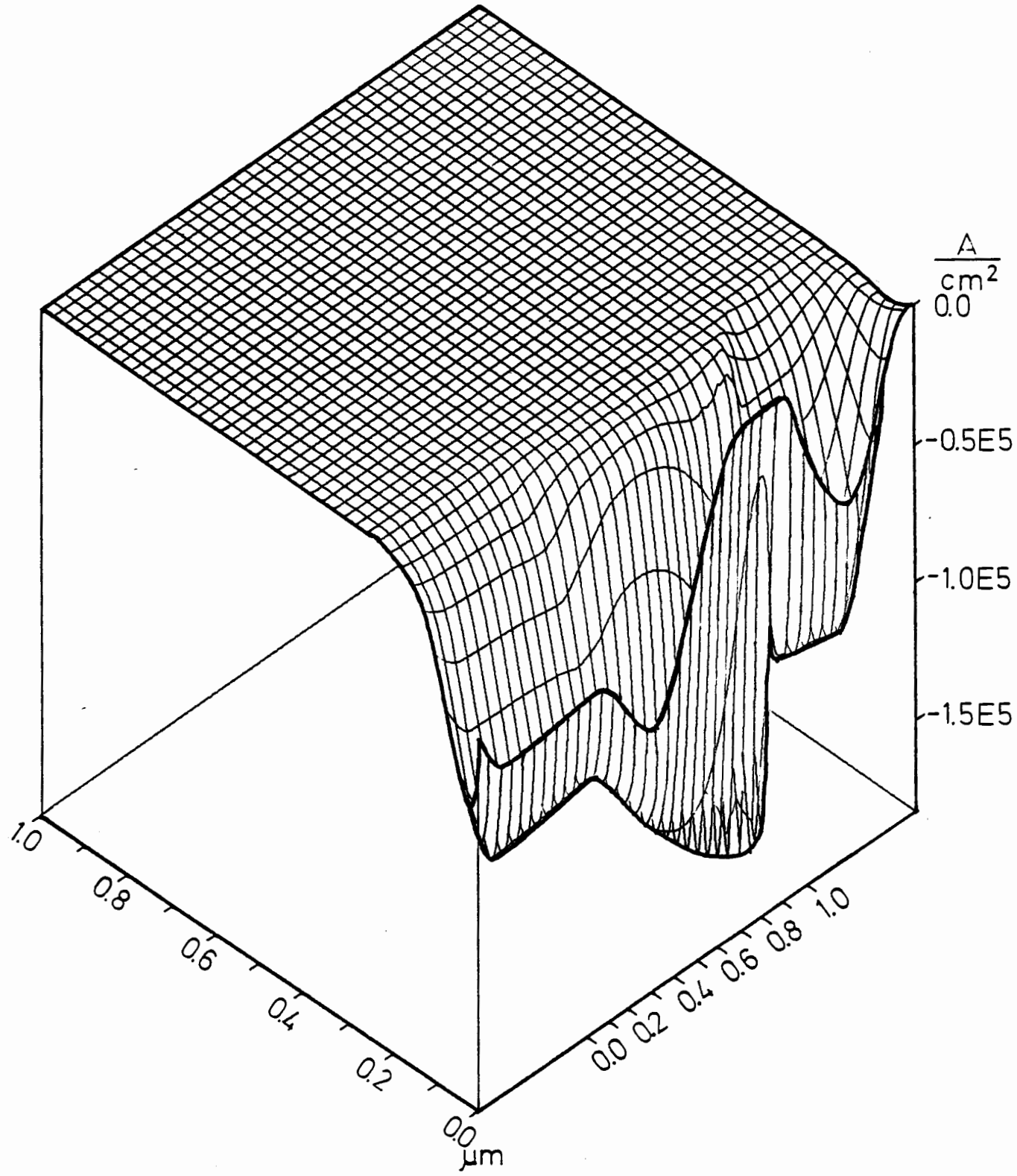


Figure 4.2-5b: The current density distribution in the load transistor.

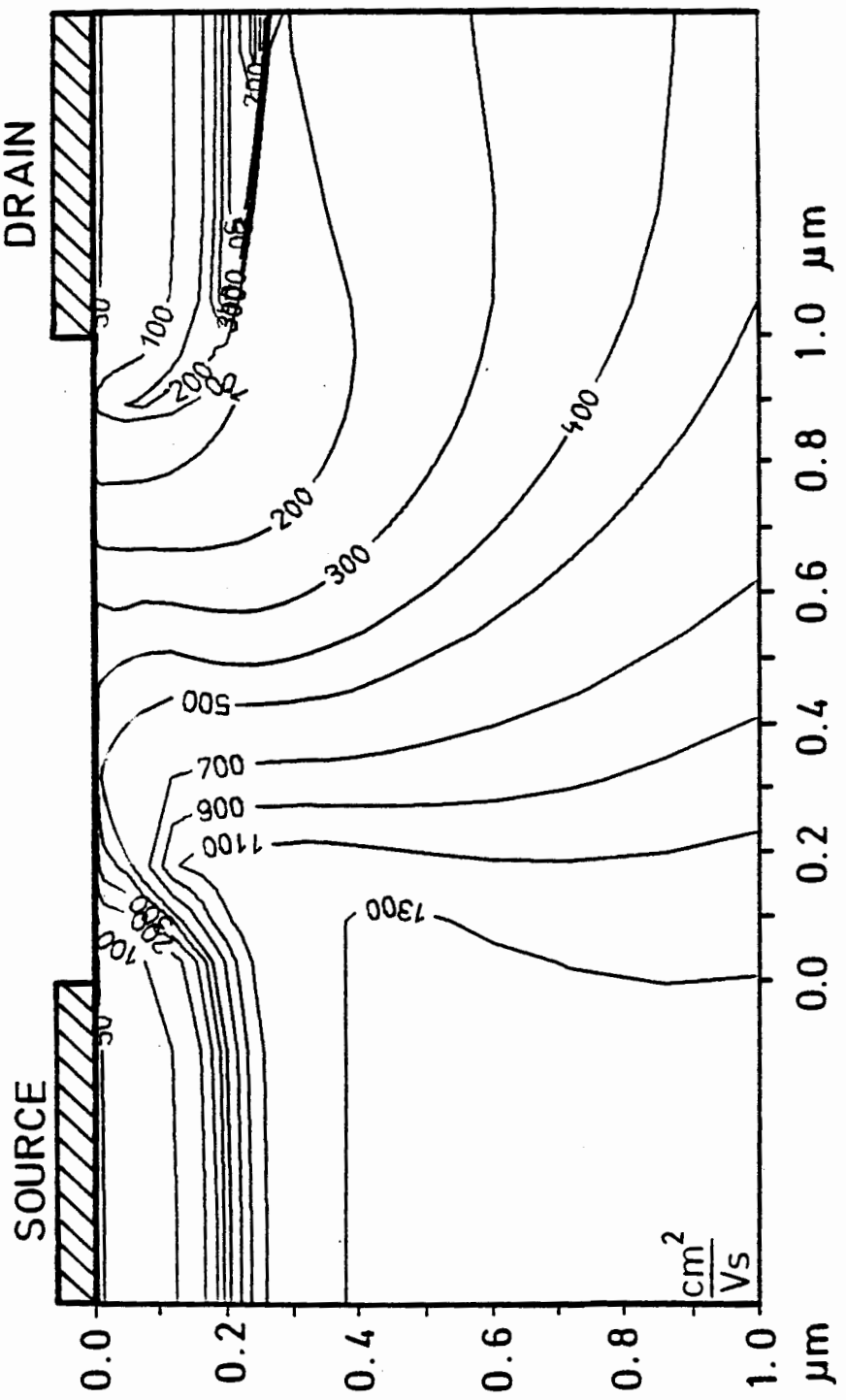


Figure 4.2-6: The mobility distribution in the load transistor.

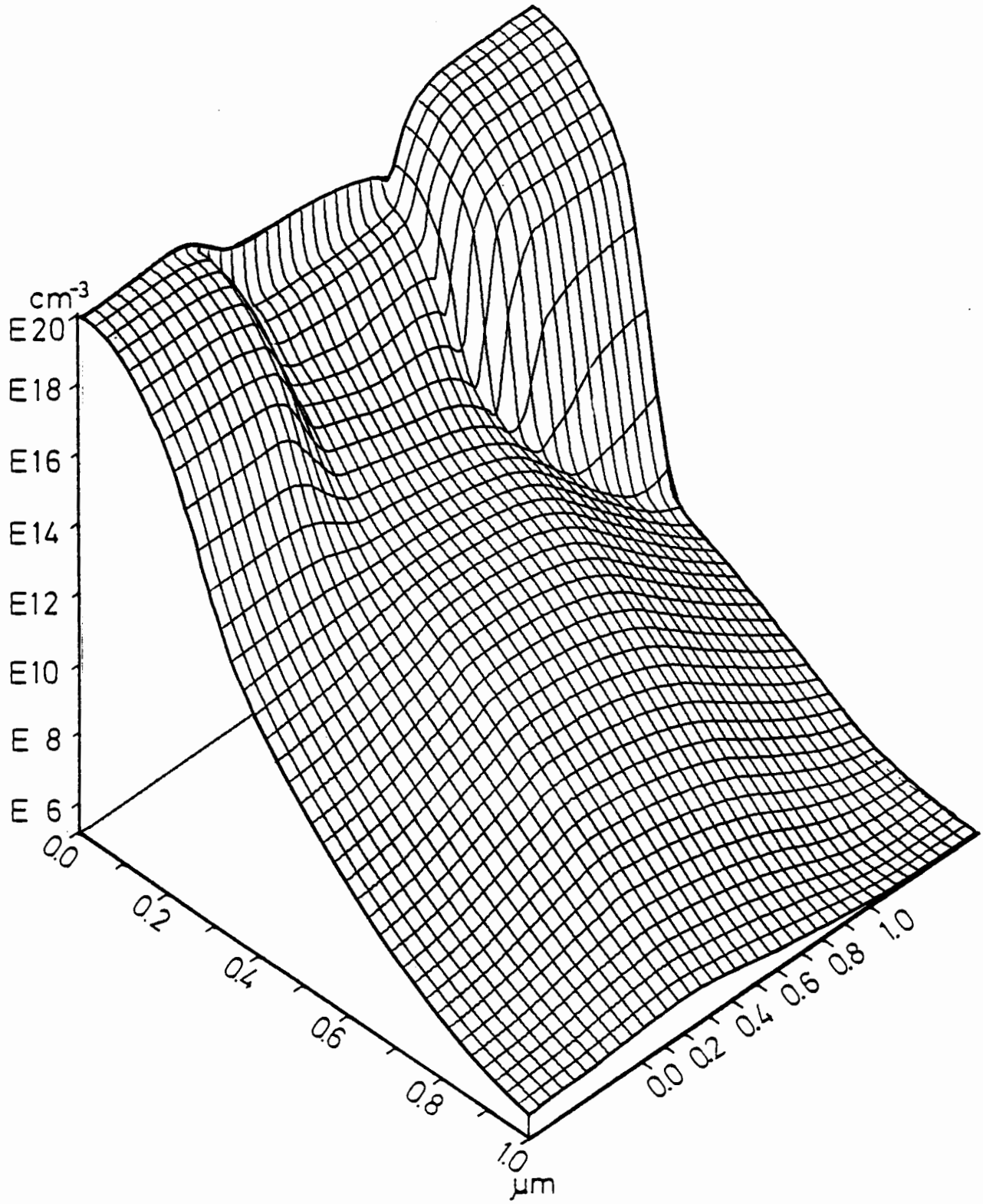


Figure 4.2-7a: The electron distribution in the load transistor.

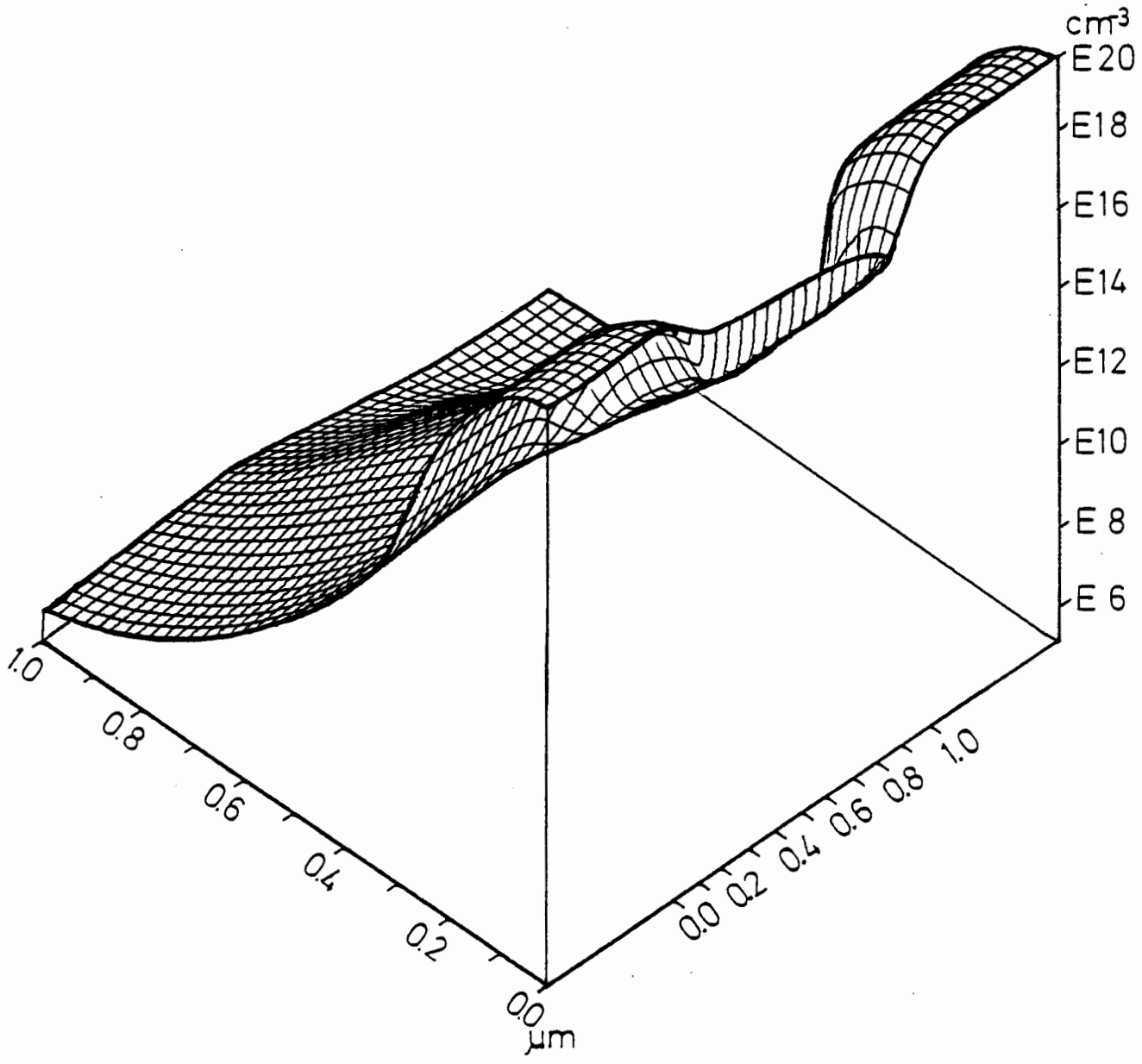


Figure 4.2-7b: The electron distribution in the load transistor.

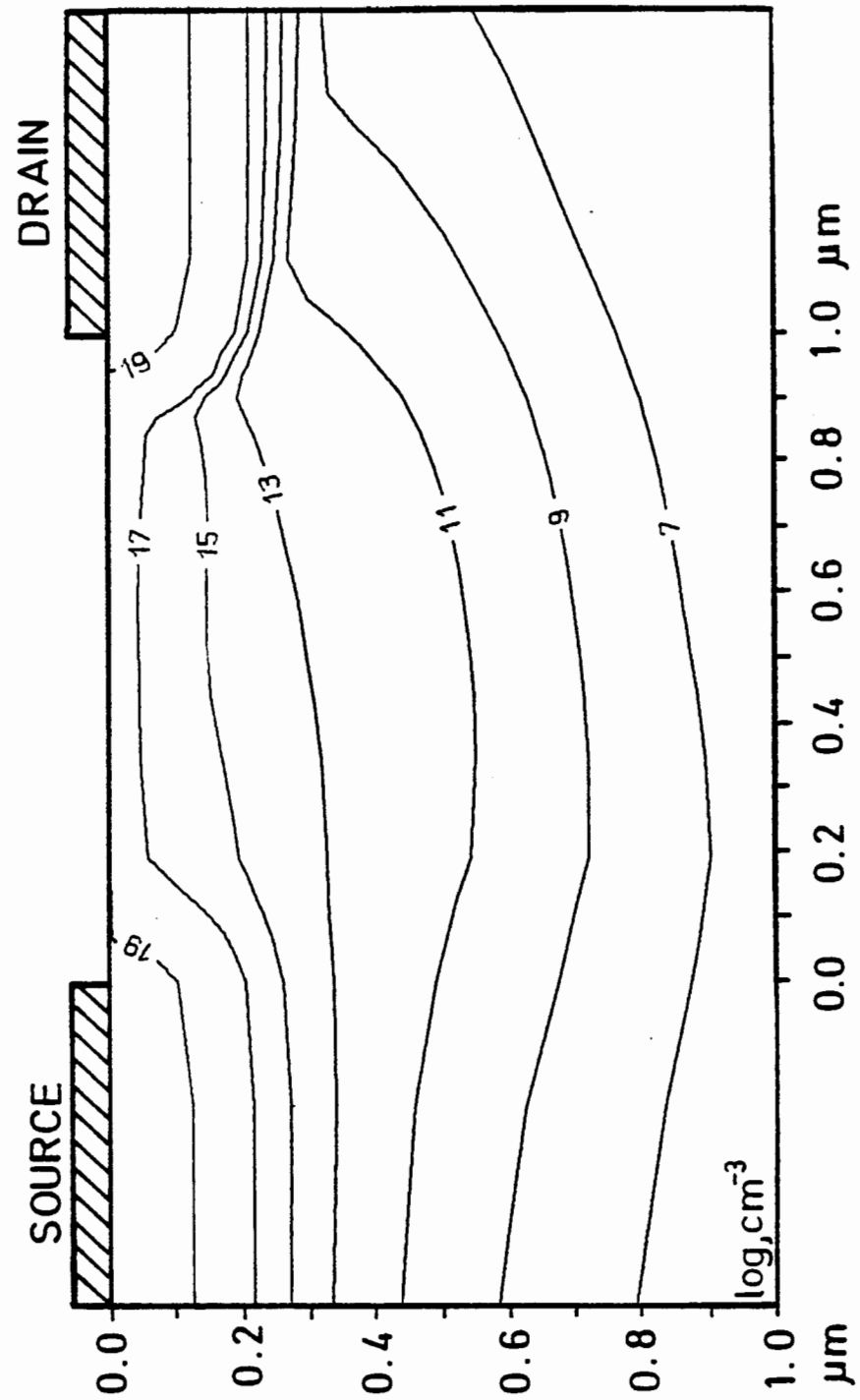


Figure 4.2-7c: The electron distribution in the load transistor.

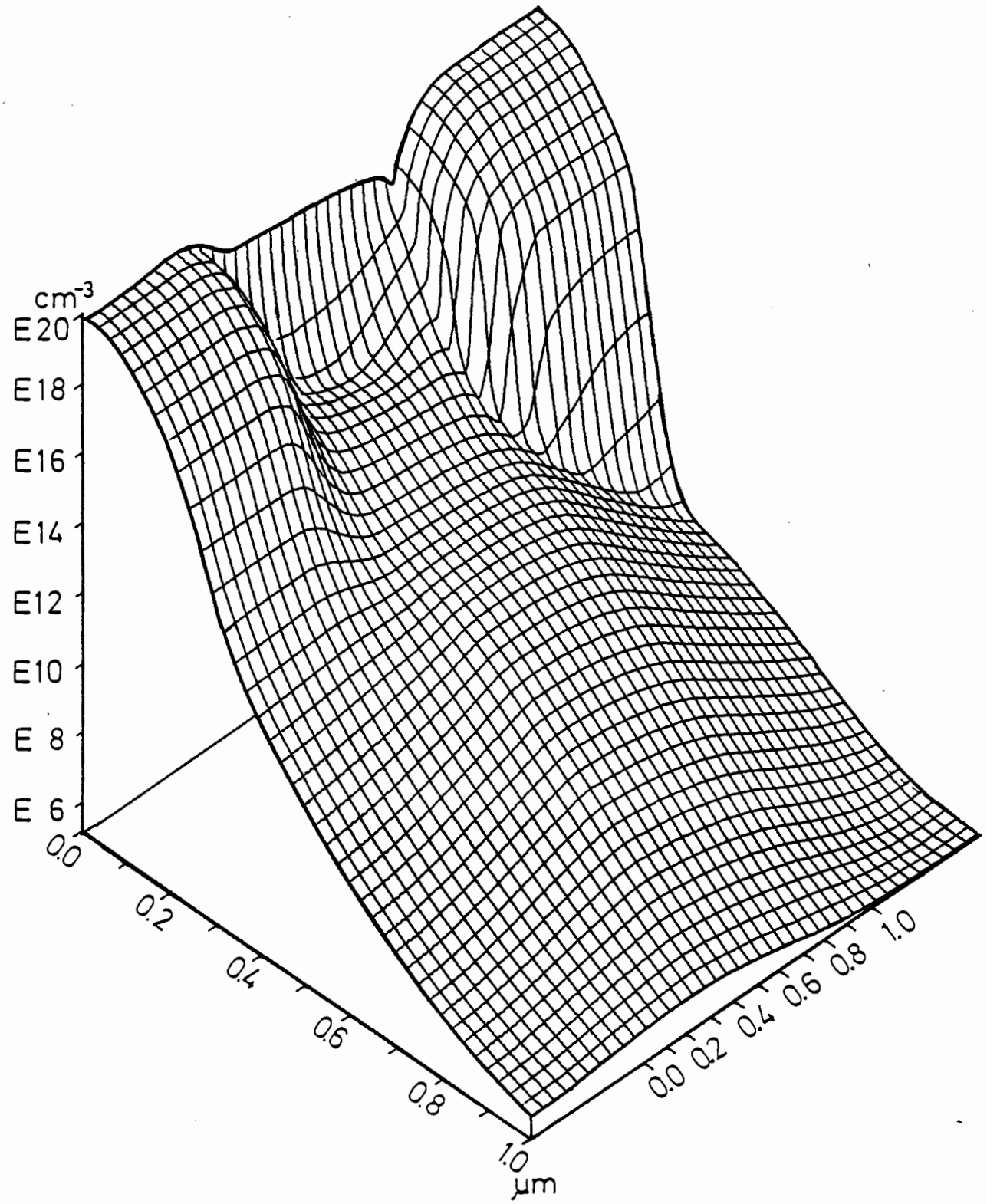


Figure 4.2-8a: The electron distribution in the drive transistor.

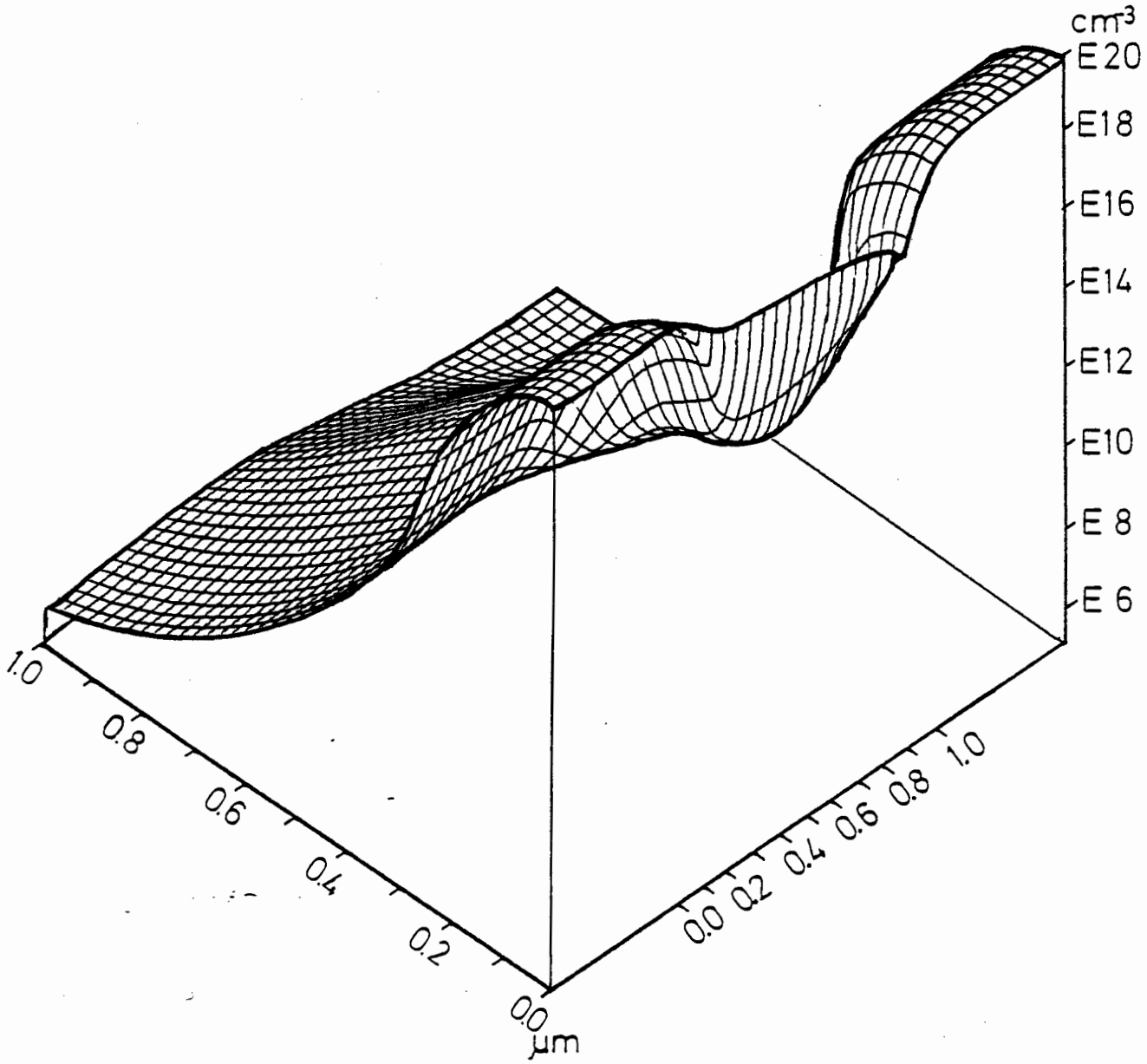


Figure 4.2-8b: The electron distribution in the drive transistor.

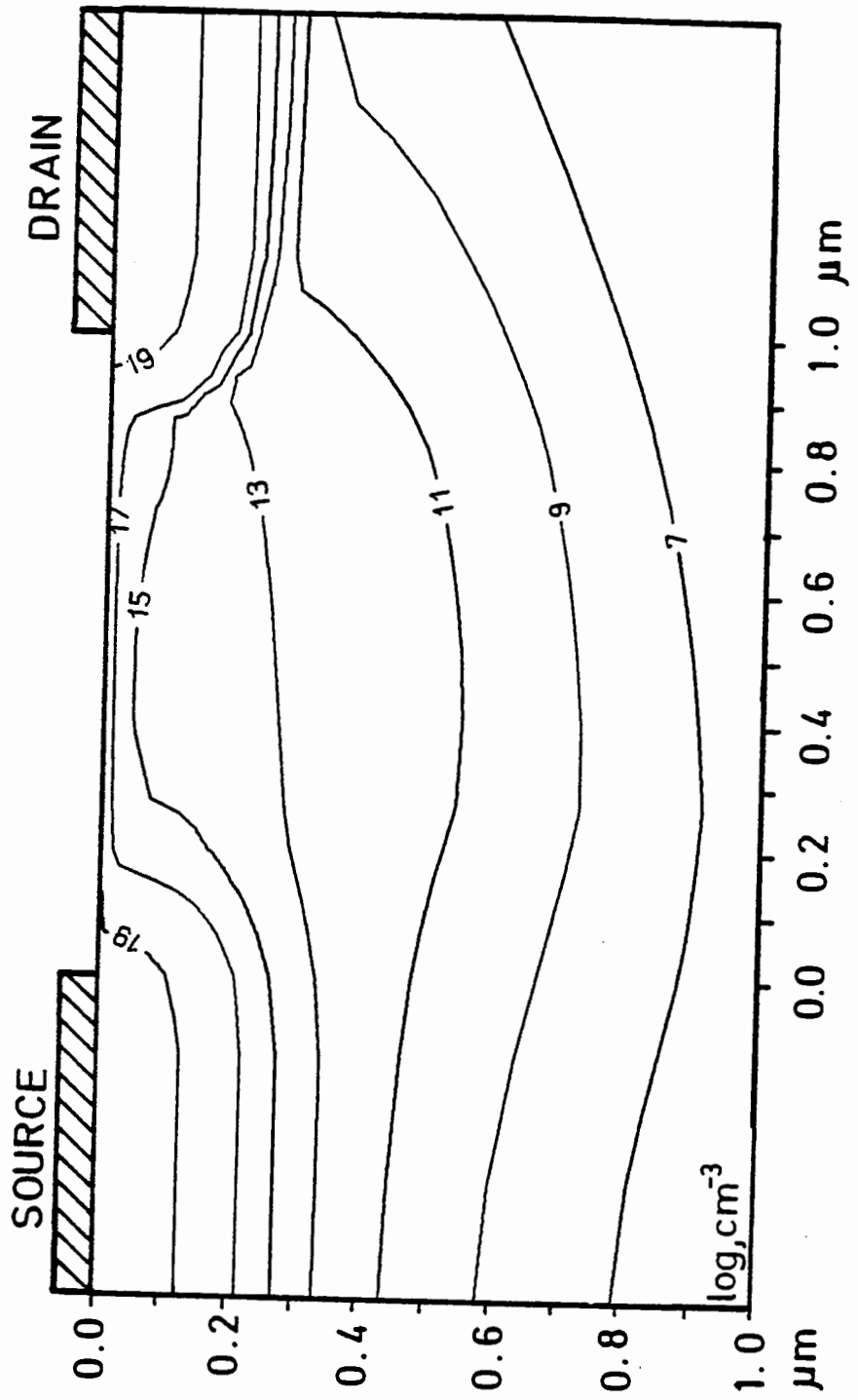


Figure 4.2-8c: The electron distribution in the drive transistor.

Figure 4.2-9 shows the characteristic curves of the drive transistor and the dashed line shows the characteristic of the load transistor. The gate and drain voltages of the load transistor are always equal for the selected circuit, and as already discussed the negative threshold voltage permits the load transistor to always operate in the ohmic region. It should be especially mentioned that the source potential is floating which has the same effect as an equivalent substrate voltage. A notable feature of the load transistor characteristic is that it is completely ohmic. This is because of a compensation effect due to the influence of the drain voltage and the substrate voltage upon the threshold voltage of the load transistor.

Figure 4.2-10 shows only the characteristic transfer curve of the inverter, which was obtained point wise from the intersections of the characteristic curves of the drive and load transistors. The low level lies at 0.2 volts; the high level is naturally at the supply voltage of 3 volts because the load transistor produces no voltage drop with no current flow. The signal to noise ratio of the low level is (at 0.2 volts) very small. This is a well known problem with miniaturized digital circuits. The reason for this lies mainly in the low threshold voltage of the drive transistor. The signal to noise ratio of the high level (at 1.36 volts) is surely satisfactory. The slope of the transfer function (with a value of -2.5) is to be expected with submicrometer logic. This could without doubt be improved by clever design of the load element. This example documents in essence, that in the near future, enormous integration densities of flawless digital circuits is to be expected by way of submicrometer technology.

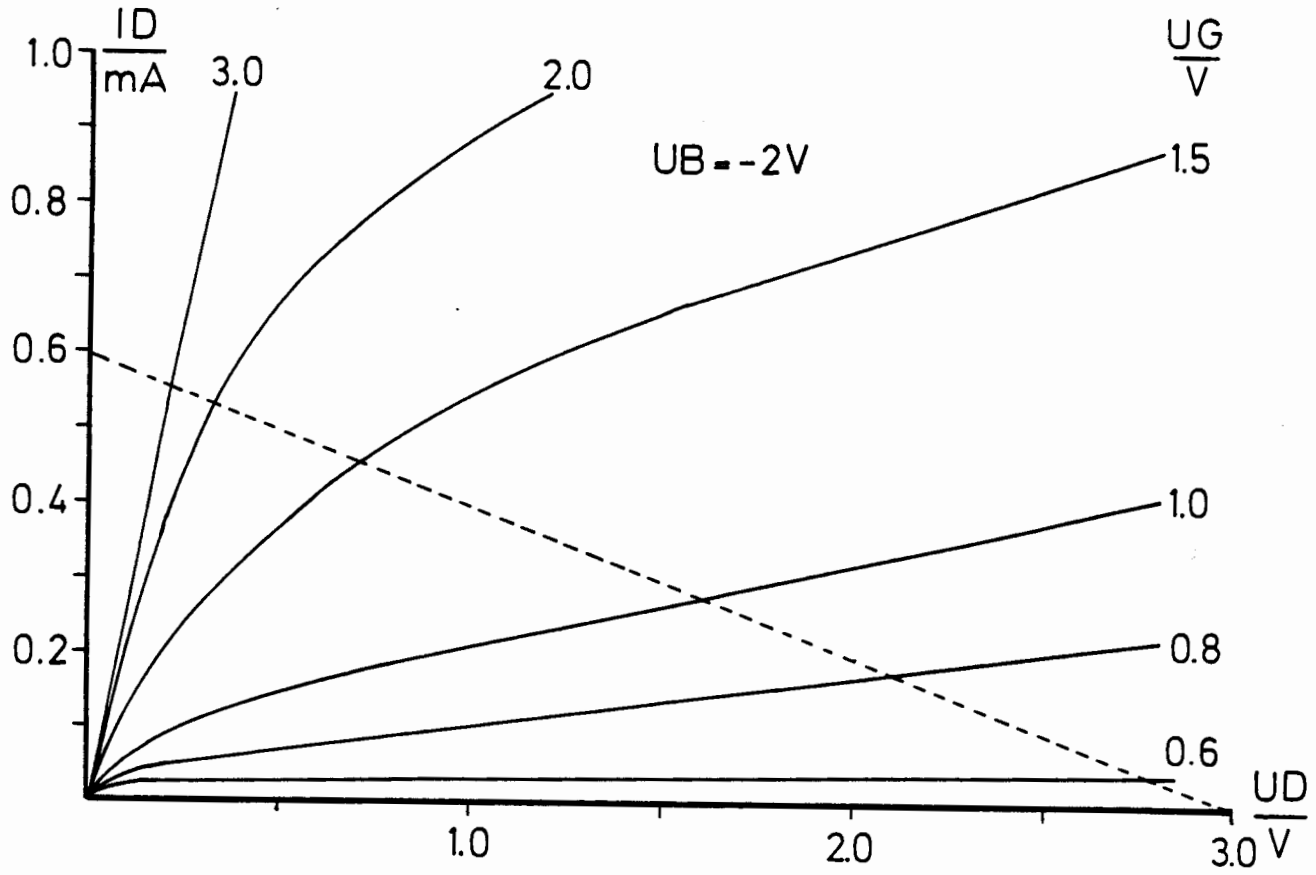


Figure 4.2-9: Drive transistor characteristic curves and the load transistor characteristic.

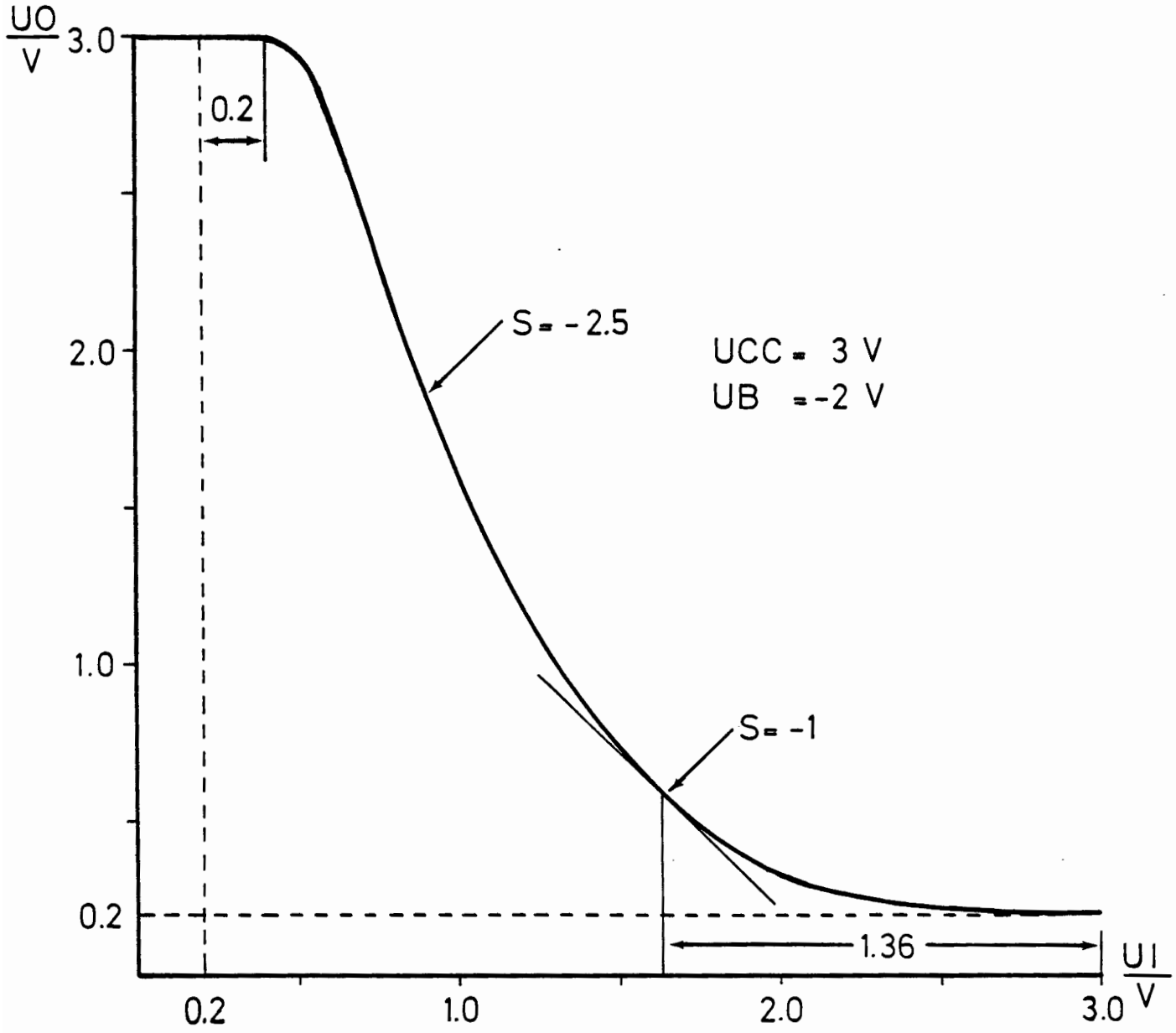


Figure 4.2-10: The transfer function of an inverter.

4.3 Process Sensitivity

Very large scale integration (VLSI) requires, without a doubt, the miniaturization of the individuual transistor elements. By reducing only the geometrical dimensions one is confronted with significant problems with respect to the electrical behavior of the transistors. One must scale all parameters together with the geometrical dimensions according to known rules, ("scaling" /38/ and /88/). In general, lower voltage, higher doping, more shallow p-n junctions and thinner oxide will be used, in order to obtain the desired behavior for the miniature transistor. For reduction of the channel length down to two micrometers, all adjustable parameters can, at the present, be controlled satisfactorily by the relevant technological steps (implantation, diffusion and oxidation). When the channel length is reduced further controllability is lost, which has been experimentally confirmed by researchers worldwide and it was naturally to be expected a priori. In general reproducibility is lost with increasing miniaturization. Especially the difficulty of controlling the parameters of nearby transistors in the same integrated circuit, which should show identical behavior, increases considerably.

In order to obtain more understanding of these conditions a process sensitivity analysis of certain transistor parameters was carried out with MINIMOS. This section will discuss in detail the sensitivity of the threshold voltage, which is one of the most interesting transistor parameters for circuit designers. It will further be attempted to find a practical limit for the miniaturization of transistors in the technology described here, a well established, yet modern MOS process. It is understood that the analysis of the threshold voltage presented here is only an example of an analysis strategy, which is usable on the remaining device parameters and technology.

4.3.1 The Transistor Which Was Analysed

The transistors investigated were made in a process which was developed for channel lengths of two micrometers. Donor doped polysilicon was used as the gate material. Arsenic was implanted for the highly doped source/drain regions in order to obtain shallow p-n junctions with a steep gradient. A double channel implantation was used in order to adjust the threshold voltage

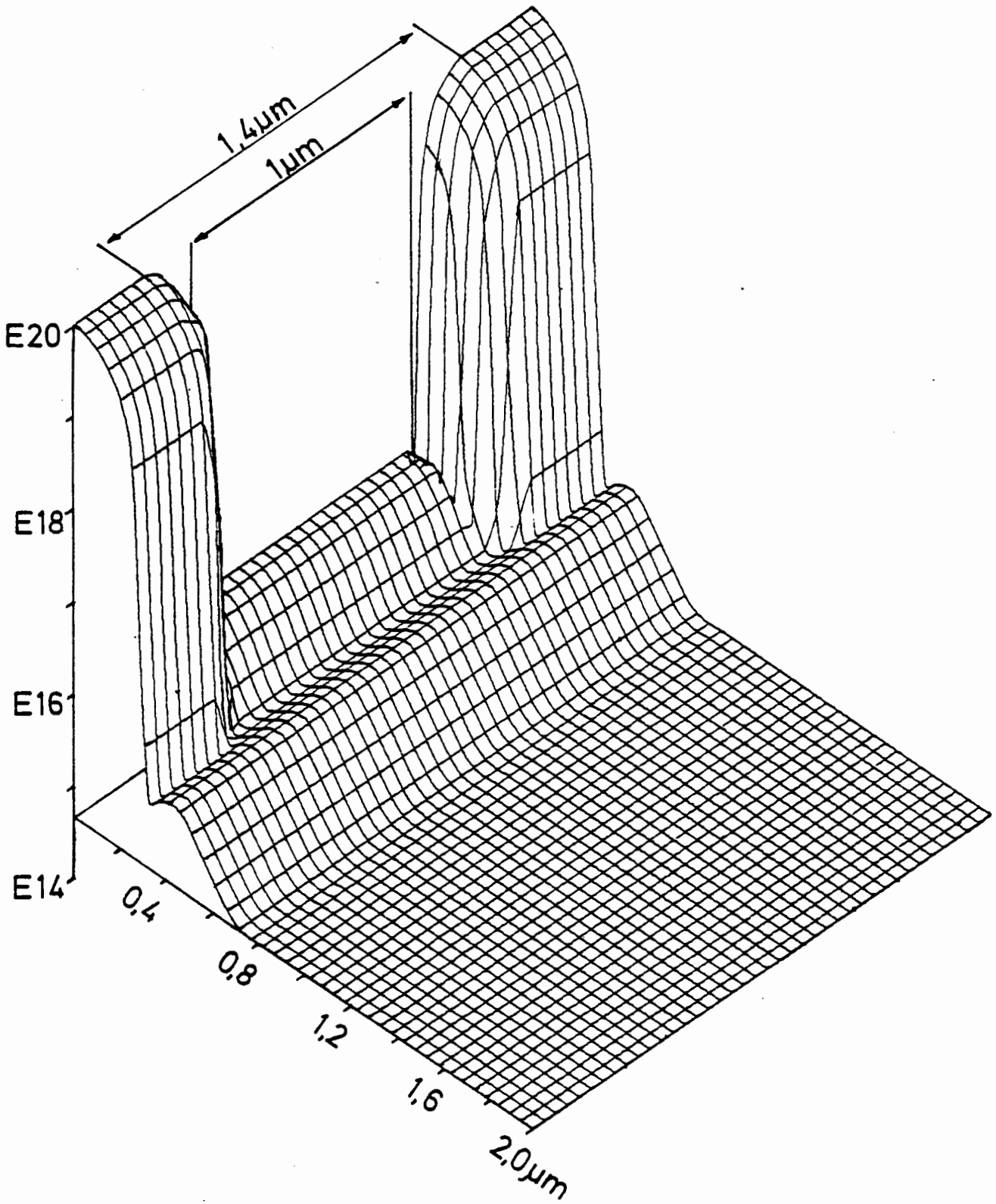


Figure 4.3-1: Doping profile of the transistor which was analysed.

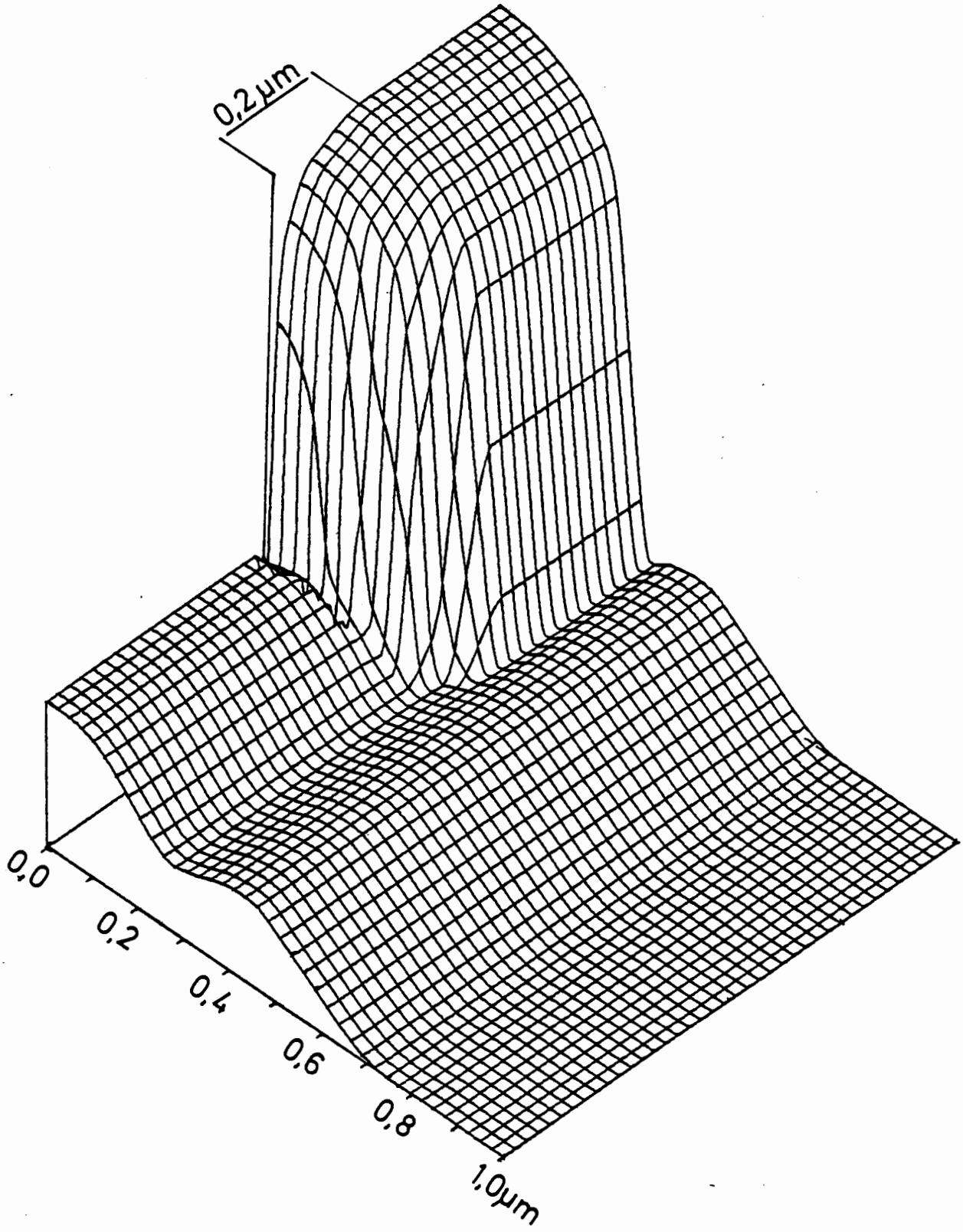


Figure 4.3-2: Enlarged detail of figure 4.3-1.

as well as to suppress "punch through".

Figure 4.3-1 shows a 3-D representation of the doping profile for a one micrometer long transistor on a logarithmic scale. It is to be noted that the scale on the lateral and vertical axes of this figure are the same. At the rear on the left is the highly doped source region and at the rear on the right is the highly doped drain region. Both channel implantations can be seen between the source and drain. The shallow implantation was carried out with a dose of $3 \cdot 10^{11}$ cm^{-2} and an energy of 35 keV and the deep implantation was carried out with a dose of 10^{11} cm^{-2} and an energy of 160 keV. Naturally, the dopant used in both cases was boron. The oxide thickness which was about 50 nanometers is not shown in this figure.

Figure 4.3-2 shows a blowup of the right rear quadrant of figure 4.3-1, in order to be able to observe greater detail. A p-n junction depth of 320 nm and an underdiffusion of about 200 nm was obtained for this process.

The process sketched here - as was described in the beginning - was developed for a two micrometer long transistor and one wants to retain as much as possible from a well established process in the development of a new process, one asks himself first and foremost the question: How will the devices made in this technology behave after a reduction in channel length?

4.3.2 The definition of threshold voltage

In order to investigate the behavior of the threshold voltage one must first give it an adequate definition. The most common definitions are based on an extrapolation of a tangent to the drain current. All of these methods are relatively inexact and the threshold voltage cannot be directly realized from all of them. The following definition was chosen because of the above reasons: The threshold voltage is that gate voltage at which the transistor sinks 0.1 microampere times the channel width per channel length. The channel length is defined as the distance between the metalurgical p-n junctions of the source-channel and drain-channel. With this definition it is insured that no threshold voltage shift versus channel length occurs for long channel transistors, which makes it possible to quantitatively measure the influence

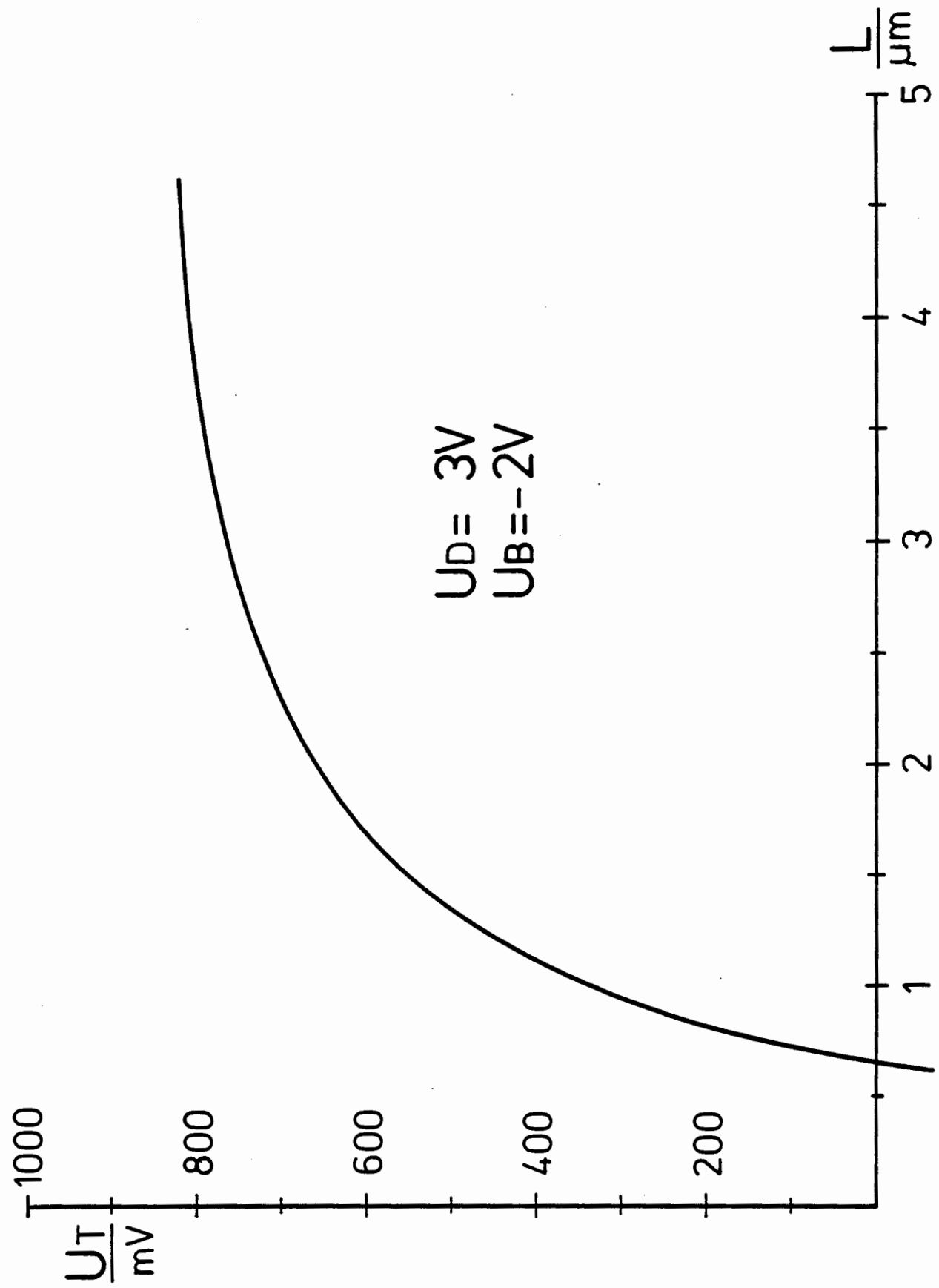


Figure 4.3-3: Threshold voltage versus channel length.

of short channel effects. Also, with the above definition, the experimental measurement of the threshold voltage is not a problem. At this point it should be mentioned that the drain voltage and the substrate voltage are not explicit parameters in the threshold voltage definition. One must obtain the dependence upon these parameters through the use of characteristic curves (threshold voltage versus drain voltage and threshold voltage versus substrate voltage).

Of course the definition chosen here is arbitrary, just as arbitrary as any other definition, and one can therefore argue about the application of the quantitative value of the selected constant (0.1 μA times channel width per channel length). Devices with a steep subthreshold characteristic - only such devices are of practical interest - produce useful results by all definitions of threshold voltage, and for devices with shallow subthreshold characteristics. A definition of threshold voltage is irrelevant.

Figure 4.3-3 shows the threshold voltage versus channel length for the transistors which were investigated. An operating point of 3 volts on the drain and -2 volts on the substrate has been chosen as a fair tradeoff for the comparison of different channel lengths. In order to avoid confusion all of the following figures will also refer to this same operating point. In figure 4.3-3 one observes the well known decrease of threshold voltage with smaller channel length. This first becomes critical as the channel length becomes less than one micrometer.

4.3.3 Sensitivities

Usually in articles about short channel MOS transistors a comparison between theoretical curves and selected experimental points is presented. Some authors also report statistical measurements (e.g. /37/), but only one publication /144/ deals explicitly with the sensitivity of an electrical transistor parameter, and the threshold voltage was indeed studied in that publication. With respect to the inherent dependence of most properties on the dispersion of geometry and technology, it seems to be a necessity to analyse and present these dependencies directly.

The computer program MINIMOS developed here is well suited for such an

analysis. One simply needs the physical model parameters for the program as for example the constants in the mobility formulation. For a transistor in a given technology it is necessary to have agreement within a small percent between simulated and measured data over a selected interval. This can be done with noncritical transistors with relatively long channels because the measured characteristics should deviate only minimally for inaccuracies in geometry and technology. In order to obtain the sensitivity one must simply vary a technological or geometrical parameter of a small transistor in the region of its nominal value and discretely differentiate to obtain the desired results, e.g., the threshold voltage. This parameter variation must certainly be done within a small range because the validity of linearization which is presupposed with the whole strategy has to be ensured. On the other hand, it is necessary to have a sufficiently large range of parameter variation to avoid cancellation errors in the (numerical) differentiation.

The above sketched procedure is not usable experimentally, because the parameter variation in a small region of high sensitivity is not reproducible. If one could exactly vary the desired parameter the possibility of experimental analysis seems questionable because the manufacturing cost would be very large and a tremendous amount of time would be required. With a good simulation program it is easy to calculate the partial derivatives of transistor parameters with respect to any desired technological or geometrical parameter by way of the above described procedure.

Figure 4.3-4 shows the partial derivative of the threshold voltage with respect to channel length versus channel length; that is, the sensitivity of the threshold voltage on the tolerances of channel length. If one assumes a transistor with an effective channel length of one micrometer with a tolerance of ten percent, which is only 100 nanometers, one can read from figure 4.3-4 a ± 60 millivolt spread in the threshold voltage. This spread is so large that a great number of integrated circuits would no longer operate properly.

Figure 4.3-5 shows the relative sensitivity of the threshold voltage on the tolerance of the gate oxide thickness versus the channel length. This dependence decreases with shorter channel length which at first glance is probably not expected. This is due to the decreasing influence of the bulk

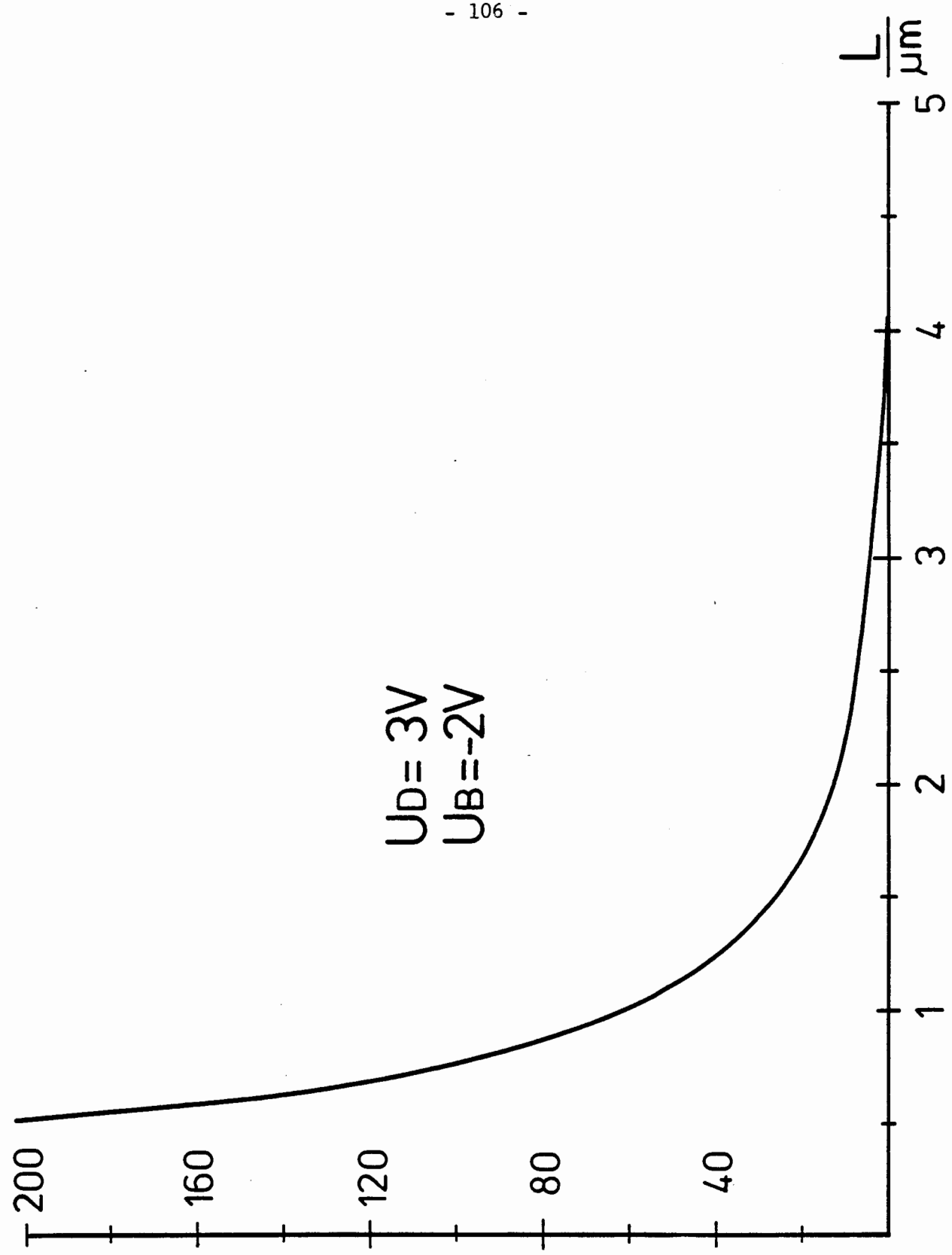


Figure 4.3-4: Sensitivity on channel length tolerances.

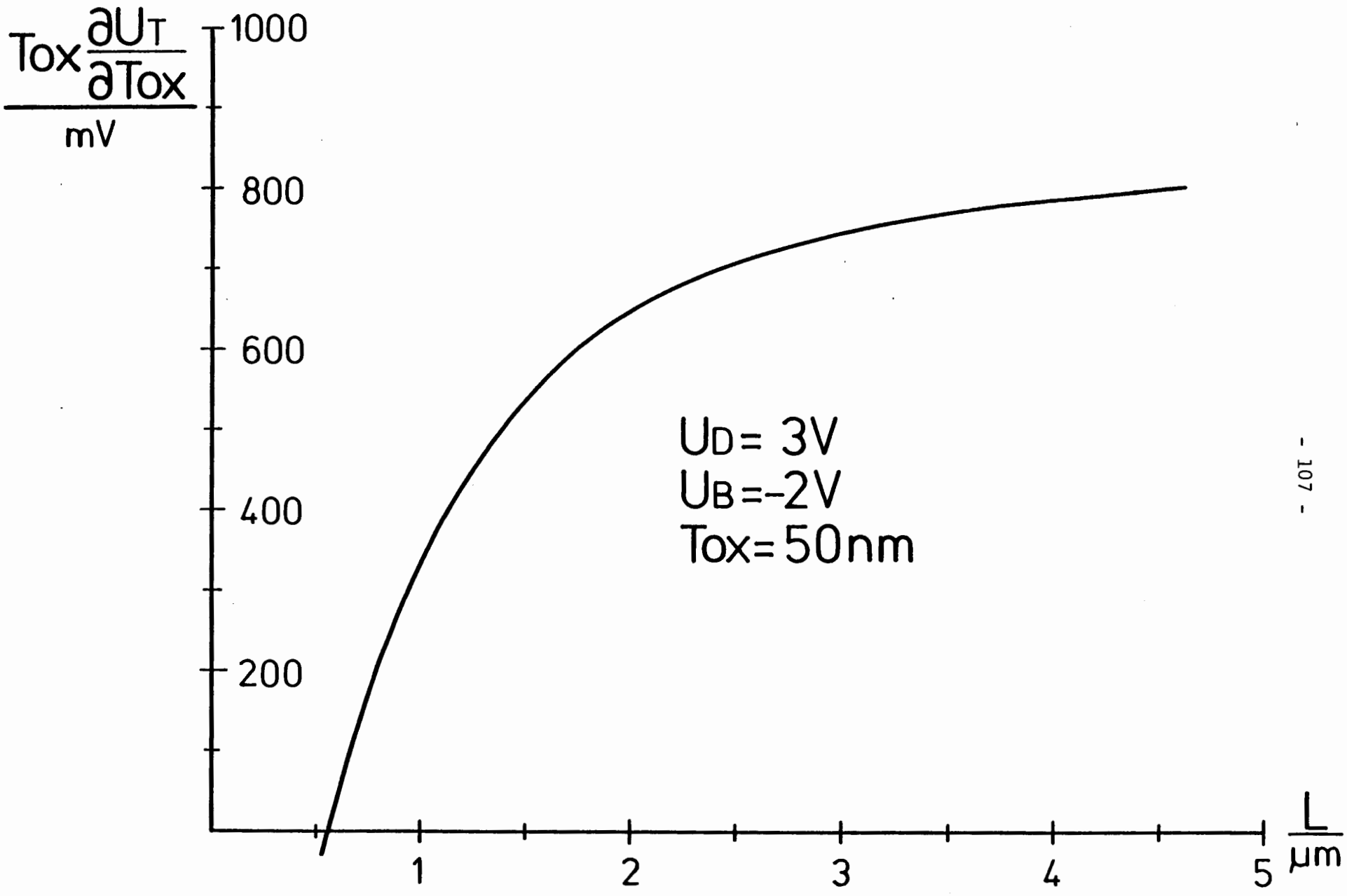


Figure 4.3-5: Sensitivity on oxide thickness tolerances.

charge with shrinking channel length. It is possibly noticeable that this figure is qualitatively similar to the figure of the threshold voltage versus channel length (figure 4.3-3). This fact is easy to understand by analytical reasoning. One must simply remember that in the first order approximation the threshold voltage is proportional to the reciprocal of the oxide capacitance per channel charge and the oxide capacitance is proportional to the reciprocal of the oxide thickness. The charge at the interface (Q_{ss}) is in general very small in comparison to the channel charge and will therefore not be considered in these qualitative calculations.

$$UT \triangleq \phi_{MS} + 2\cdot\phi_F - (Q_{ss}+Q_b)/C_{ox}$$

(without short channel effects)

$$C_{ox} = \varepsilon_{ox}/T_{ox}$$

$$\partial UT/\partial T_{ox} \triangleq -(Q_{ss}+Q_b)/\varepsilon_{ox}$$

$$UT \triangleq (\partial UT/\partial T_{ox})\cdot T_{ox} + \text{const.}$$

The decrease in the threshold voltage with decreasing channel length (not included in the above formulation) basically lies physically in the reduction of the channel charge by way of the space charge regions of the source-substrate and drain-substrate diodes, and yet one spreads the channel charge over the oxide, therefore the decrease in the sensitivity of the threshold voltage with respect to oxide thickness tolerance appears plausible from this point of view. As a developer of transistors one should not be delighted by this decrease in dependence upon oxide thickness, which means at the same time, as was mentioned in the last sentence, that the controllability of the transistor by way of the gate is decreased.

Figure 4.3-6 shows the sensitivity of threshold voltage on junction depth tolerances versus channel length. A one micrometer transistor with an uncertainty of ten percent has an uncertainty in the threshold voltage of ±40 millivolts. Again as one can read from this figure, there exists, in general, no influence on the threshold voltage of long channel transistors by the p-n

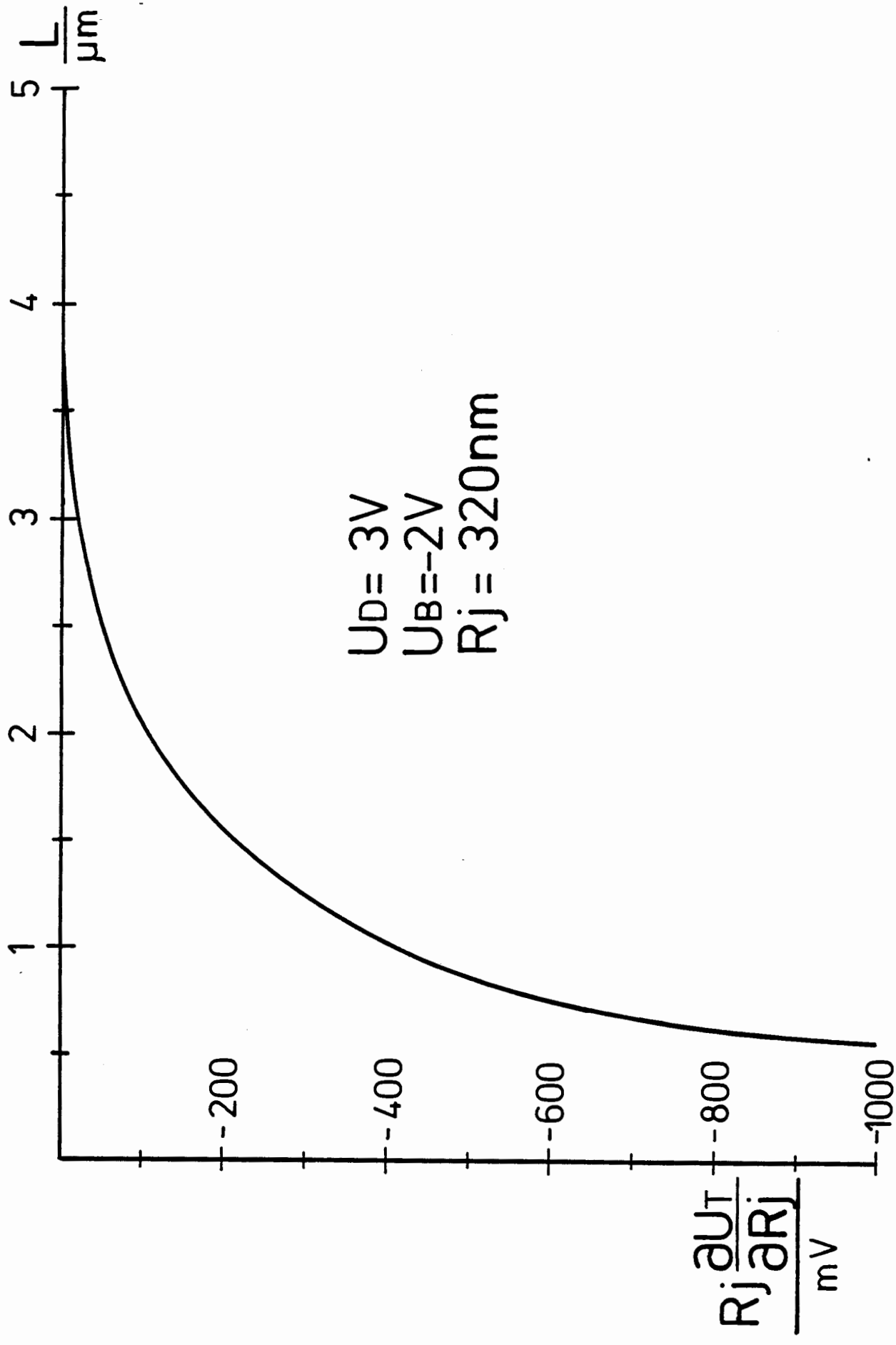


Figure 4.3-6: Sensitivity on junction depth tolerances.

junction depths. Physically this effect can be explained relatively easily: At a given critical channel length the source-substrate and drain-substrate reverse-biased diodes are in a position to influence the charge in the channel. This influence is simply the short channel effect (see also /133/). If the transistor is so short that both space charge regions almost touch each other, the channel charge will not only be reduced, but much more, the barrier of the source-channel will be shorted and at about the depth of the p-n junction a conduction channel will be created, the "punch through" channel. Both effects, insofar as will be discussed here, produce a reduction in controllability of the transistor by way of the gate and result in a dependence on the p-n junction depth. On the one hand, the threshold voltage will be shifted in the negative direction and additionally, what is more undesirable, the steepness of the subthreshold characteristics will be reduced. A well controlled p-n junction depth proves to be a required necessity for the manufacturing of short channel transistors.

Figure 4.3-7 shows the sensitivity of threshold voltage on drain bias variation. A 300 millivolt change in drain voltage, that is, ten percent of the applied bias, results in about 30 millivolts change in the threshold voltage for this operating point. Physically this effect is based on the influence of the space charge region of the drain-substrate diode on the channel charge. This influence has already been discussed in the explanation of figure 4.3-6, the sensitivity on junction depth tolerances. It naturally makes little difference, whether the space charge region was a result of junction depth or uncertainty in the drain voltage, the qualitative similarities of both sensitivities are easy to understand.

The sensitivity on drain bias variation could be measured with sufficient experimental effort. However in the case of short channel devices only the nominal values of the process and technological parameters are known. On the basis of the expected dispersion of these parameters the value of the measurements would be degraded. One could only make statistical measurements and extract statistical bars on each measured point. This procedure would be experimentally expensive and the expected statistical bar would be so large, that one would not be able to draw any convincing conclusions from the measurements.

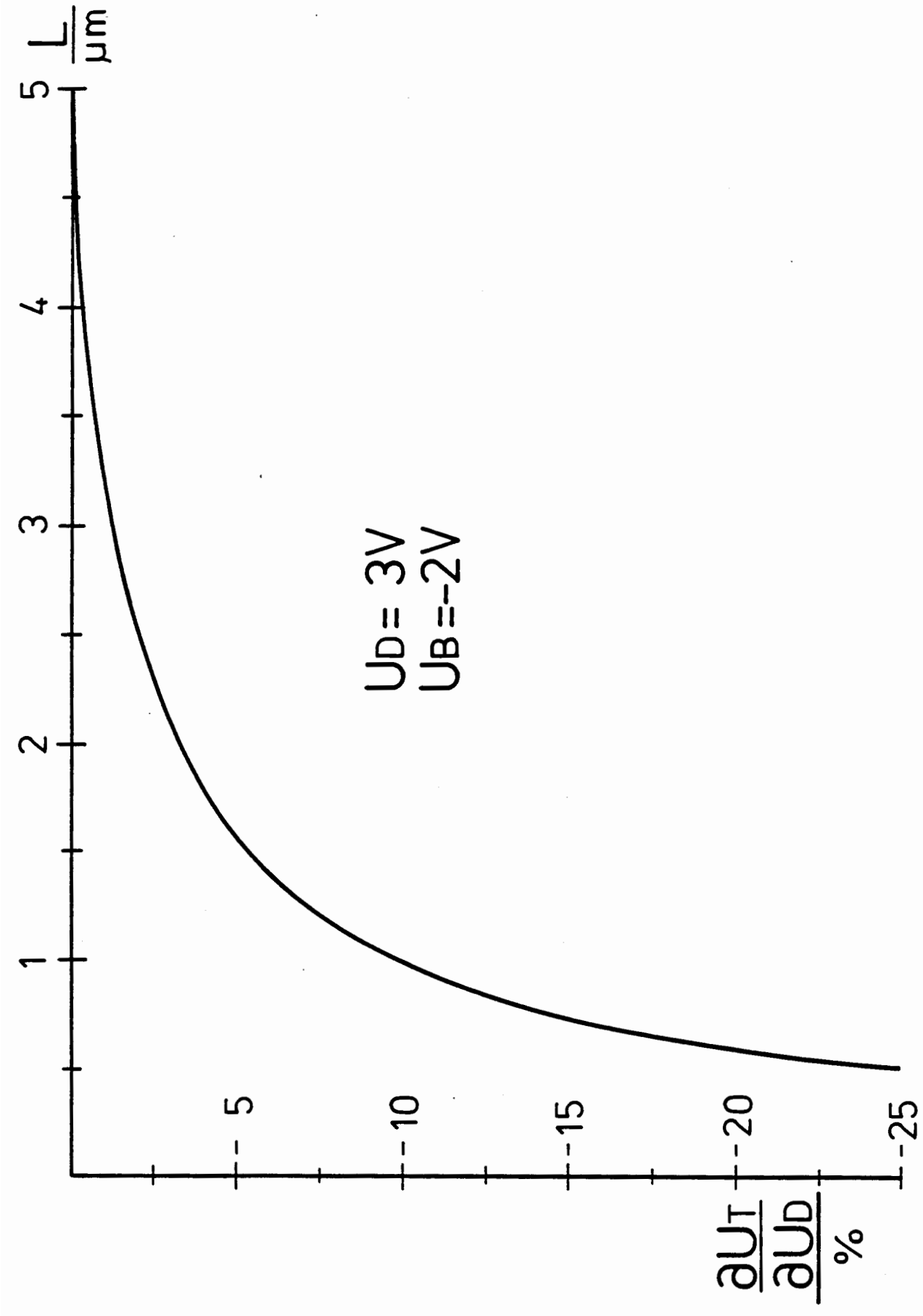


Figure 4.3-7: Sensitivity on drain bias variation.

Figure 4.3-8 shows the sensitivity of threshold voltage on bulk bias variation. A 200 millivolt change in the bulk bias, ten percent, results in a threshold shift of about 11 millivolts, which is usually noncritical. For the long channel transistor one can easily estimate this sensitivity analytically. It is namely:

$$\partial UT/\partial UB \triangleq -1/C_{ox} \cdot \partial Q_b/\partial UB$$

with

$$Q_b \triangleq q \cdot N_b \cdot y_c$$

y_c is the space charge zone width under the channel
N_b is the substrate doping

For the partial derivative with respect to substrate bias only y_c, the space charge zone width is nonconstant. Therefore this means that:

$$\partial UT/\partial UB \triangleq -(T_{ox}/y_c) \cdot (\varepsilon_{si}/\varepsilon_{ox})$$

By estimating y_c at about two micrometers and the quotient of ε_{si} to ε_{ox} at about three one obtains a value of about 7.5 percent for the sensitivity of bulk bias for long channel transistors, which the exact two dimensional calculations confirm. Although as has already been said, this sensitivity is relatively small, in practical cases all the sensitivities are summed together, so that one should not completely forget this effect.

An interesting detail of this figure is the fact that the sensitivity decreases first with shrinking channel length and at a critical length begins to increase rapidly. One can interpret this behavior as a superposition of short-channel effect and "punch-through". The short channel effect produces a decreasing tendency of the sensitivity, which means that the substrate voltage loses control over the channel charge. In the case of "punch-through" the sensitivity again increases, in that the substrate voltage causes a field that definitely influences the action of the drain.

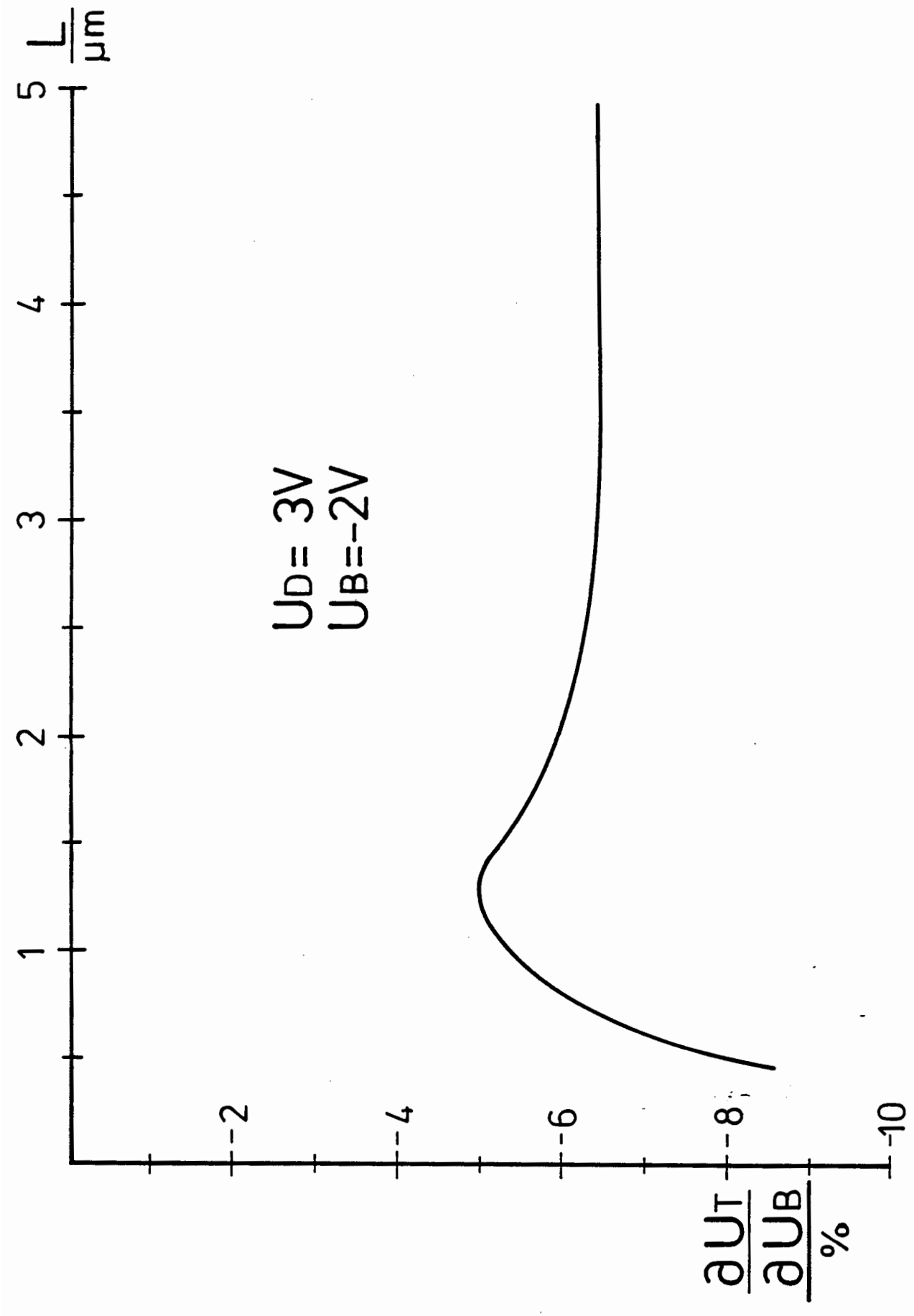


Figure 4.3-8: Sensitivity on bulk bias variation.

Figure 4.3-9 shows the influence of fluctuations in the energy of the steep channel implantation on the threshold voltage. Qualitatively one again observes the superposition of the "punch-through" effect and the short channel effect. The absolute value of this particular sensitivity is low due to the fact that the depletion region below the channel covers the entire implanted region.

Figure 4.3-10 shows the sensitivity of the threshold voltage on the uncertainties of the dose of the steep channel implantation. This figure is, as was to be expected, very similar to figure 4.3-9, the sensitivity on uncertainties in implantation energy. For the long transistor it is possible to estimate this sensitivity in a simple manner. The implantation dose is simply placed in the channel charge:

$$Q_b \triangleq q \cdot (N_b \cdot y_c + \text{Dose})$$

The threshold voltage sensitivity is calculated to be:

$$\partial UT/\partial \text{Dose} \triangleq -1/C_{ox} \cdot \partial Q_b/\partial \text{Dose}$$
$$= q/C_{ox} = 23\ \text{mV}/10^{10}\text{cm}^{-2}$$

Figure 4.3-11 shows the temperature coefficient of threshold voltage on the transistors investigated. One also distinctly observes the previously discussed behavior, namely, the superposition of short-channel effects and "punch-through" effects. The absolute value of this sensitivity lies around a value of about -1 mV/°K. The absolute value as well as the qualitative behavior has been verified by experiments /129/. For the long-channel transistor one can simply calculate an approximate value for the temperature coefficient. The principle temperature dependent term in the simple threshold voltage formula is namely the fermi level. Hence:

$$\partial UT/\partial T \triangleq 2 \cdot \partial \phi_F/\partial T \triangleq (U_g - 2 \cdot \phi_F)/T = -1.5\ \text{mV/K}$$

The two dimensional calculations give a value of exactly -1 mV/°K which can be taken from figure 4.3-11.

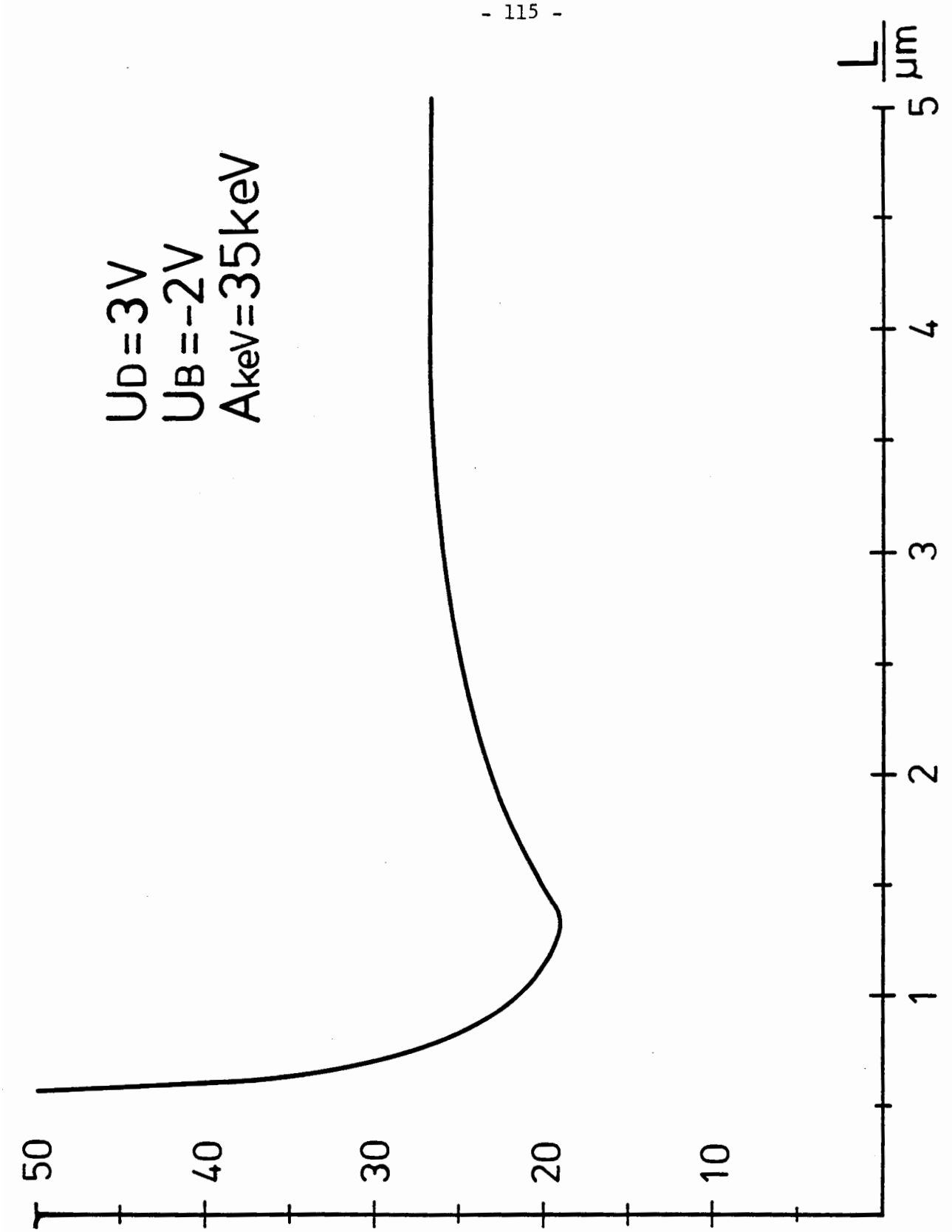


Figure 4.3-9: Sensitivity on implantation energy tolerances.

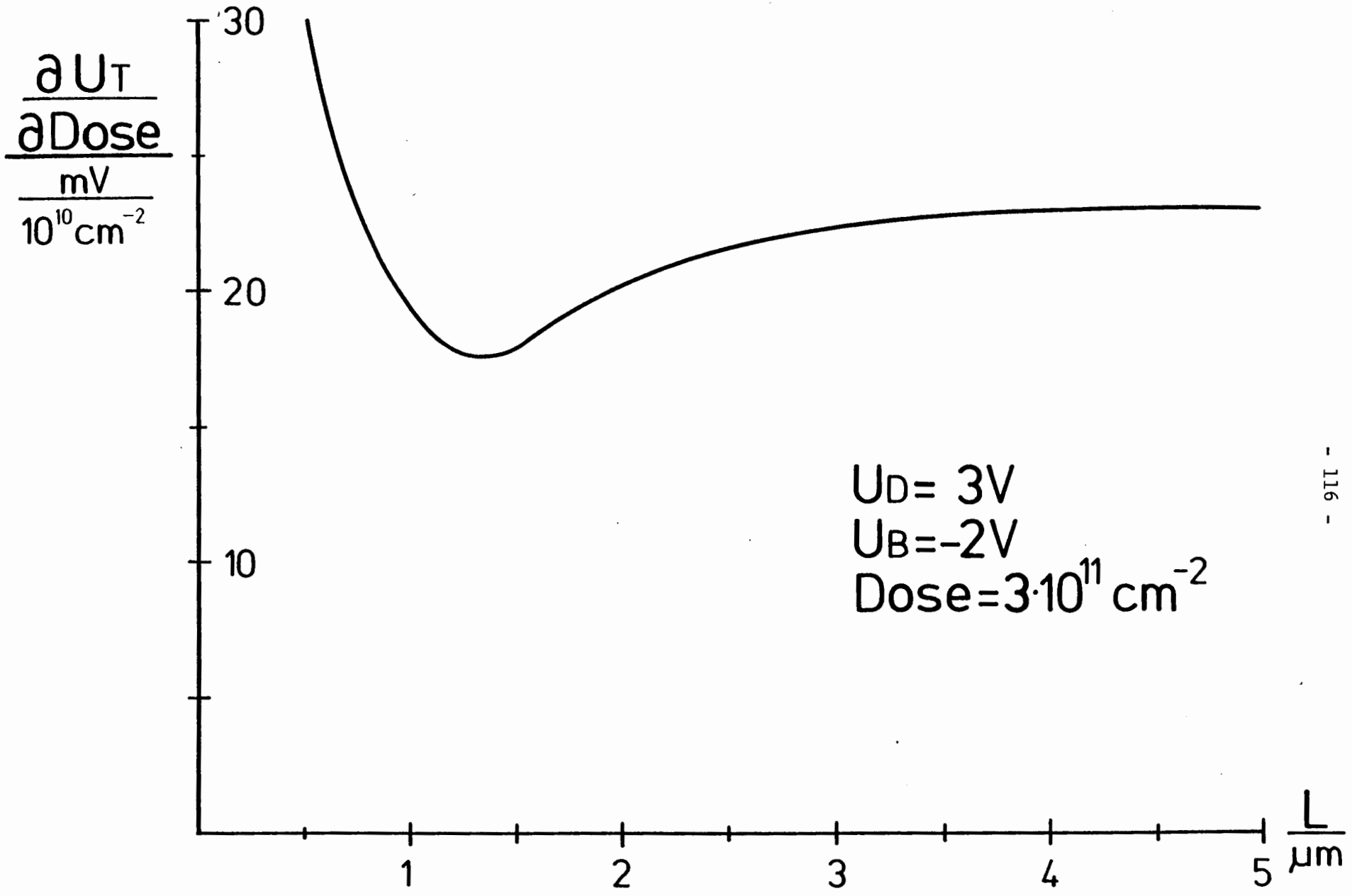


Figure 4.3-10: Sensitivity on implantation dose tolerances.

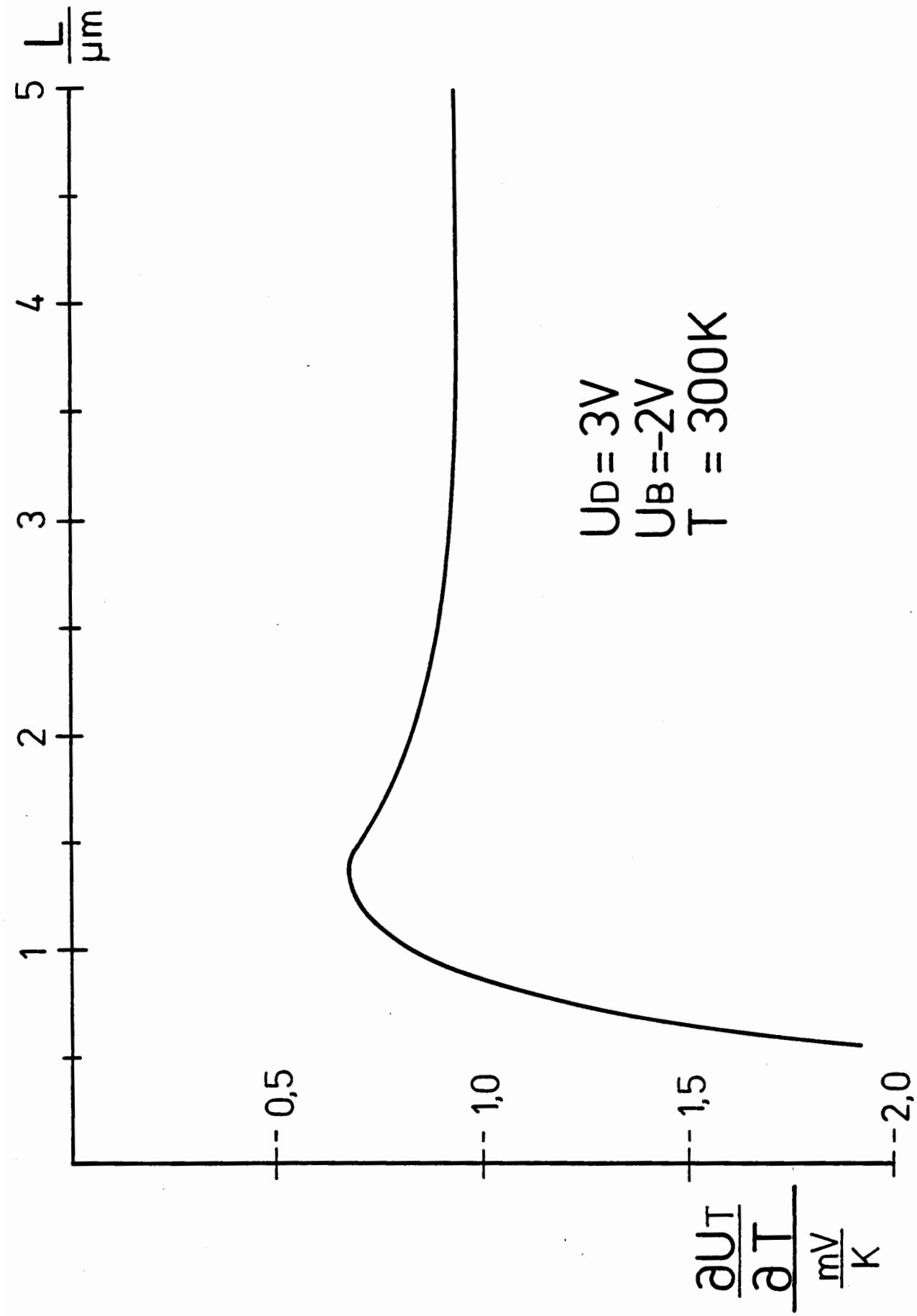


Figure 4.3-11: Sensitivity on temperature variation.

Figure 4.3-12 shows the subthreshold characteristics with L/W scaled for the transistors investigated with channel length as the parameter. One observes that the shift of these characteristic curves due to short channel effects is very small down to a channel length of about two micrometers. A more noticeable fact is that the slopes of these characteristic curves first begin to deteriorate with channel lengths under one micrometer, which verifies the occurrence of ("punch-through") which has already been discussed many times with respect to previous figures. The fact that some of those figures show a local extremum with a calculated channel length of about 1.3 micrometers can also not be completely explained here with this figure. As has already been pointed out, the occurrence of an increased short channel effect together with the collapse of the subthreshold slope ("punch-through") is responsible for the limits to any further miniaturization of the transistors in the technology investigated here.

4.3.4 Global Sensitivity

The partial derivatives discussed in the last subsection denote isolated sensitivities on a certain set of parameters. One can obtain from these results which parameters are the most critical with respect to the development of a device. A global sensitivity which describes the sum of all detailed sensitivities is of great interest for many reasons. Such a sensitivity should be related to a specific technology and its applications and should indicate the limit of channel length reduction. In order to calculate such a global sensitivity typical ranges of deviations of design parameters have to be specified.

Figure 4.3-13 shows an example of quantitative values of these ranges for a one micrometer technology. The channel length uncertainty is shown to be relatively small with an absolute value of 100 nanometers. One must remember that for a one micrometer technology this already amounts to ten percent.

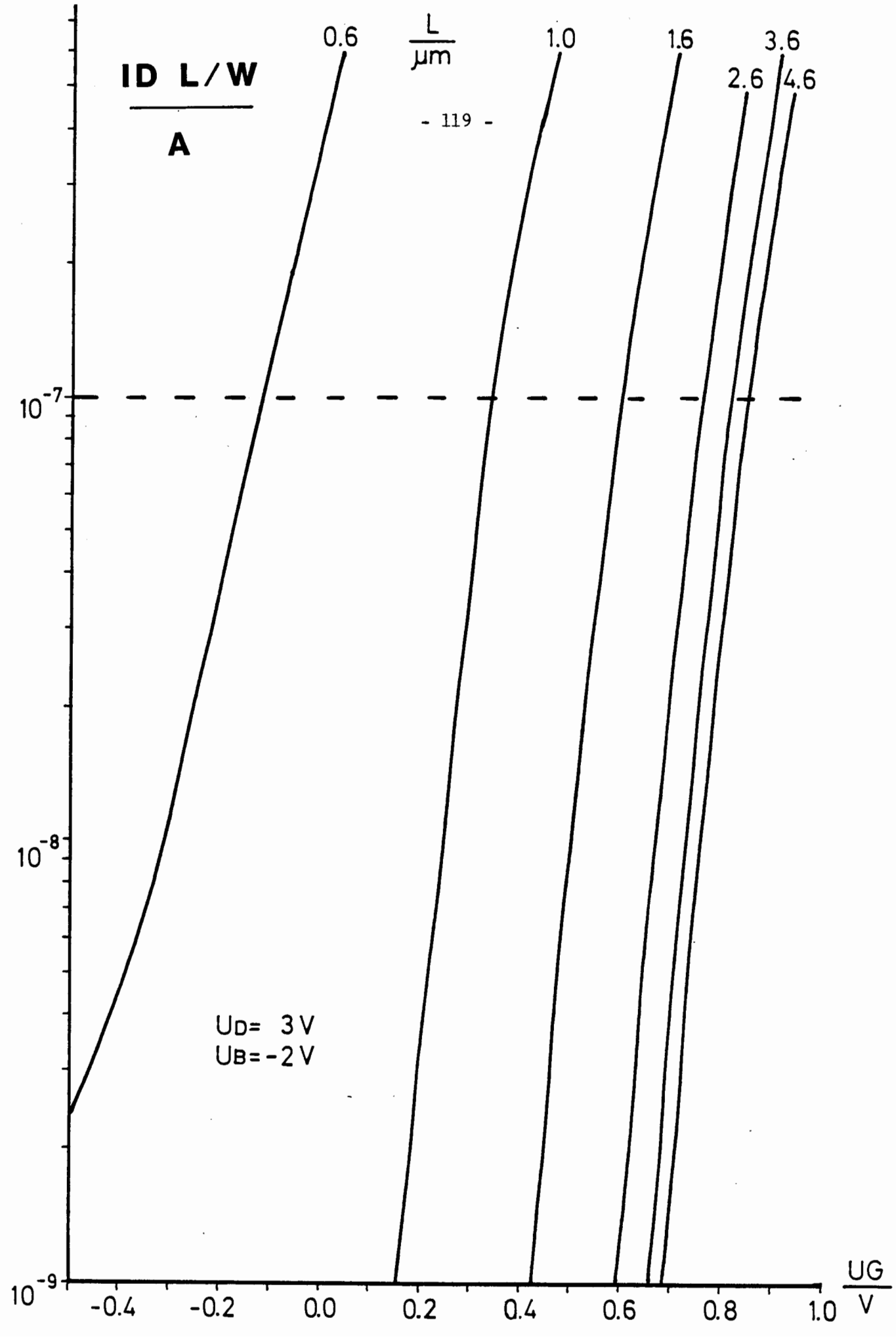


Figure 4.3-12: Scaled subthreshold characteristics.

Parameter	X	$\lvert\Delta X\rvert$	%
L		100 nm	
TOX	50 nm	2.5 nm	5
RJ	320 nm	32 nm	10
UD	3 V	150 mV	5
UB	-2 V	100 mV	5
AKEV	35 keV	0.7 keV	2
DOSE	$3 \cdot 10^{11}$ cm^{-2}	$6 \cdot 10^{9}$ cm^{-2}	2

Figure 4.3-13: Desired process operating tolerances.

Obtaining oxide thicknesses within five percent and p-n junction depths within ten percent should be realizable in a good manufacturing facility. The circuit designer is expected to stabilize the bias voltages within five percent. At the present, to control the implantation parameters within two percent seems to be an extremely high demand. With modern equipment this is also realizable.

Figure 4.3-14 shows the global sensitivity which was calculated with the above given specifications. σ_D denotes the uncertainty of the threshold voltage with respect to neighboring devices, which should behave identically in the same integrated circuit. D stands for device. This sensitivity is given only by the channel length variation as the other parameters are commonly very homogeneous in a given circuit. σ_W and W stand for wafer and denote the uncertainty of threshold voltage of transistors, which should behave identically, but which have been fabricated in different process runs. In this case one must use a type of Euclidean norm over all detailed sensitivities in order to obtain a global sensitivity. It it noteworthy that

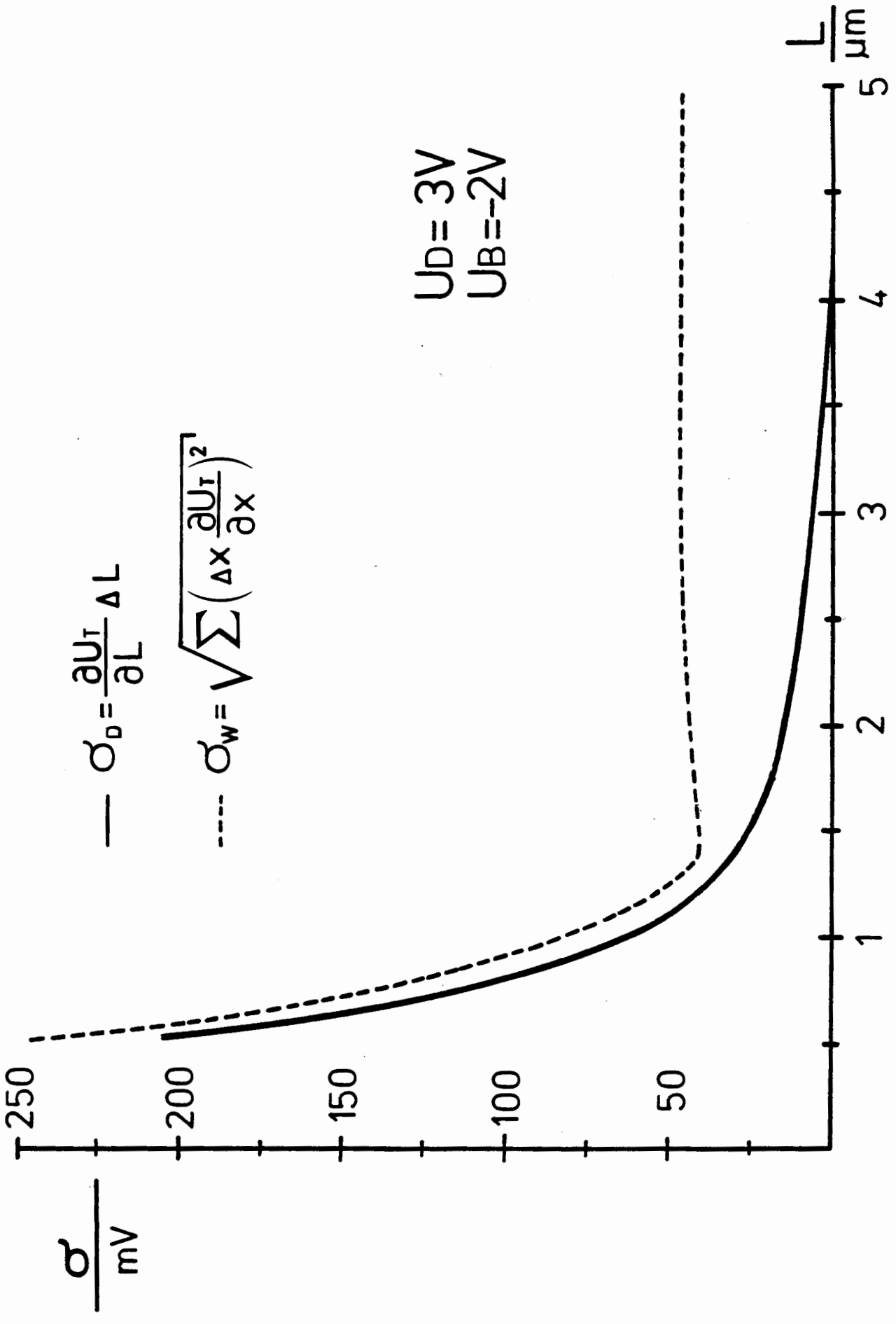


Figure 4.3-14: Reproducibility of the devices investigated versus channel length.

this curve seems to be constant down to a certain channel length at which an excellently pronounced knee is located, 1.4 micrometers, which can be interpreted as the practical limit of channel length reduction due to the threshold uncertainty. Physically this limit depends strongly on the beginning of the degradation of the subthreshold characteristics. To obtain a practical limit for miniaturization purely from subthreshold characteristics is a difficult problem in that the degradation begins very slowly and therefore it will be too late when the limit is first recognized. It should here then be stated, what most device developers unanimously maintain and should not be taken trivially, that the subthreshold characteristics represent one of the most important indicators of the quality of a transistor design.

5. Conclusion and Outlook

In the scope of this work a static, two dimensional model for miniaturized planar MOS transistors was presented. A user oriented computer program, MINIMOS, was developed, whose two dimensional model has a physical basis. The main motivations for this work were as follows:

* To more deeply understand the behavior of modern miniaturized MOS transistors.

* To bridge the gap between technology modeling and computer aided circuit design.

* To be able to place at one's disposal a simple to manage yet very exact tool for MOS transistor simulations.

It was shown by selected examples that the above mentioned goals were satisfactorily fulfilled. Especially the knowledge of the distributions of important physical quantities such as charge carrier distributions and current densities were made possible. It has also been made possible to gain a deeper insight into the functioning of devices of this type which could not have been done in any other way.

Furthermore, the presented work provides the basis for further modeling efforts. The present work can be modified with fundamental physical models on the one hand to improve the understanding of MOS transistors, in order to be able to analyse the breakdown behavior due to impact ionization or switching behavior and on the other hand to be able to analyse other devices which can be analysed from "first principles" would prove to be interesting and necessary.

Literature Cited

/1/ Adachi T., Yoshii A., Sudo T., "Two-Dimensional Semiconductor Analysis Using Finite-Element Method", IEEE Trans. Electron Devices, Vol. ED-26, pp. 1026-1031,1979.

/2/ Adler, M.S., "A Method for Terminating Mesh Lines in Finite Difference Formulations of the Semiconductor Device Equations", Solid-State Electron., Vol. 23, pp.845-853, 1980.

/3/ Adler M.S., "A Method for Achieving and Choosing Variable Density Grids in Finite Difference Formulations and the Importance of Degeneracy and Band Gap Narrowing in Device Modeling", Proc. NASECODE I Conf., pp.3-30, 1979.

/4/ Adler M.S., Temple V.A.K., "The Dynamics of the Thyristor Turn-On Process", IEEE Trans. Electron Devices, Vol.ED-27, pp.483-494, 1980.

/5/ Antoniadis D.A., Gonzales A.G., Dutton R.W., "Boron in Near-Intrinsic (100) and (111) Silicon under Inert and Oxidizing Ambients", J. Electrochem. Soc., Vol.125, No.5, pp.813-819, 1978.

/6/ Antoniadis D.A., Hansen S., Dutton R.W., "Suprem II - a Program for IC Process Modeling and Simulation", Stanford Technical Report No.5019-2, 1978.

/7/ Barnes J.J., "A Two-Dimensional Simulation of MESFET's", Dissertation, University of Michigan, 1976.

/8/ Barnes J.J., Lomas R.J., "Finite-Element Methods in Semiconductor Device Simulation", IEEE Trans. Electron Devices, Vol.ED-24, pp.1082-1089, 1977.

/9/ Barnes J.J., Shimohigashi K., Dutton R.W., "Short-Channel MOSFET's in the Punchthrough Current Mode", IEEE Trans. Electron Devices, Vol.ED-26, pp.446-453, 1979.

/10/ Barnes J.J., Lomas R.J., "Transient 2-Dimensional Simulation of a Submicrometre Gate-Length M.E.S.F.E.T.", Electron. Lett., Vol.11, pp.519-521, 1975.

/11/ Bell D.A., "Improved Formulation of Gummel's Algorithm for solving the 2-Dimensional Current-flow Equations in Semiconductor Devices", Electron. Lett., Vol.8, No. 22, pp.536-538, 1972.

/12/ Birkhoff G., "The Numerical Solution of Elliptic Equations", SIAM, Philadelphia, 1971.

/13/ Bozler C.O., Alley G.D., "Fabrication and Numerical Simulation of the Permeable Base Transistor", IEEE Trans. Electron Devices, Vol.ED-27, pp.1128-1141, 1980.

/14/ Brown G.W., Lindsay B.W., "The Numerical Solution of Poisson's Equation for Two-Dimensional Semiconductor Devices", Solid-State Electron., Vol.19, pp.991-992, 1976.

/15/ Buturla E.M., Cotrell P.E., "Simulation of Semiconductor Transport Using Coupled and Decoupled Solution Techniques", Solid-State Electron., Vol.23, pp.331-334, 1980.

/16/ Buturla E.M., Cotrell P.E., Grossman B.M., Salsburg K.A., Lawlor M.B., McMullen C.T. "Three-Dimensional Finite Element Simulation of Semiconductor Devices", Proc. International Solid-State Circuits Conf., pp.76-77, 1980.

/17/ Canali C., Majni G., Minder R., Ottaviani G., "Electron and Hole Drift Velocity Measurements in Silicon and Their Empirical Relation to Electric Field and Temperature", IEEE Trans. Electron Devices, Vol. ED-22, pp.1045-1047, 1975.

/18/ Caughey D.M., Thomas R.E., "Carrier Mobilities in Silicon Empirically Related to Doping and Field", Proc. IEEE, Vol. 52, pp.2192-2193, 1967.

/19/ Cherednichenko D.I., Gruenberg H., Sarkar T.K., "Solution to a Diffusion Problem with Mixed Boundary Conditions.", Solid-State Electron., Vol.17, pp.315-318, 1974.

/20/ Chin D.J., Kump M.R., Lee H.G., Dutton R.W., "Process Design Using Coupled 2D Process and Device Simulators", Proc. International Electron Devices Meeting, pp.223-226, 1980.

/21/ Chryssafis A., Love W., "A Computer-Aided Analysis of One-Dimensional Thermal Transients in n-p-n Power Transistors", Solid-State Electron., Vol.22, pp.249-256, 1979.

/22/ Coe D.J., Brockman H.E., Nicholas K.H., "A comparison of Simple and Numerical Two-Dimensional Models for the Threshold Voltage of Short-Channel MOSTs", Solid-State Electron., Vol.20, pp.993-998, 1977.

/23/ Coen R.W., Muller R.S., "Velocity of Surface Carriers in Inversion Layers on Silicon", Solid-State Electgron., Vol.23, pp.35-40, 1980.

/24/ Colak S., Singer B., Stupp E., "Lateral DMOS Power Transistor Design", IEEE Electron Device Letters, Vol.EDL-1, pp. 51-53, 1980.

/25/ Control Data Corporation, "Common Memory Manager Version 1 Reference Manual", CDC, Publ.No. 60499200, 1978.

/26/ Control Data Corporation, "Cyber Loader Version 1 Reference Manual", CDC, Publ. No. 60429800, 1979.

/27/ Control Data Corporation, "Fortran Extended Version 4 Reference Manual", CDC, Publ.No. 60497800, 1978.

/28/ Control Data Corporation, "NOS/BE Version 1 Reference Manual", CDC, Publ.No. 60493800, 1979.

/29/ Control Data Corporation, "Update 1 Reference Manual", CDC Publ.No. 60449900, 1978.

/30/ Cotrell P.E., Buturla E.M., "Two-Dimensional Static and Transient Simulation of Mobile Carrier Transport in a Semiconductor", Proc. NASECODE I Conf., pp.31-64, 1979.

/31/ D'Avanzo D.C., "Modeling and Characterization of Short-Channel Double Diffused MOS Transistors", Stanford Technical Report No. G-201-6, 1980.

/32/ Dang L.M., Konaka M., "A Two-Dimensional Computer Analysis of Triode-Like Characteristics of Short-Channel MOSFET's", IEEE Trans. Electron Devices, Vol.ED-27, pp.1533-1539, 1980.

/33/ De La Moneda F.H., "Threshold Voltage from Numerical Solution of the Two-Dimensional MOS Transistor", IEEE Trans. Circuit Theory, Vol.CT-20, pp.666-673, 1973.

/34/ De Mari A., "An Accurate Numerical One-Dimensional Solution of the P-N Junction under Arbitrary Transient Conditions", Solid-State Electron., Vol.11, pp.1021-2053, 1968.

/35/ De Mari A., "An Accurate Numerical Steady-State One-Dimensional Solution of the P-N Junction", Solid-State Electron., Vol.11, pp.33-58, 1968.

/36/ Debye P.P., Conwell E.M., "Electrical Properties of N-Type Germanium", Physical Review, Vol. 93, pp.693-706, 1954.

/37/ Demoulin E., Greenfield J.A., Dutton R.W., Chatterjee P.K., Tasch A.F., "Process Statistics of Submicron MOSFET's", Proc. International Electron Devices Meeting, pp.34-37, 1979.

/38/ Dennard R.H., Gaennslen F.H., YU H.N., Rideout V.L., Bassous E., Le Blanc A.R., "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions", IEEE J. Solid-State Electron., Vol.SC-9, pp.256-268, 1974.

/39/ Dongarra J.J., Moler C.B., Bunch J.R., Stewart G.W., "LINPACK", SIAM, Philadelphia, 1979.

/40/ Dubock P., "D.C. Numerical Model for Arbitrarily biased Bipolar Transistors in Two Dimensions", Electron. Lett., Vol. 6, pp.53-55, 1970.

/41/ Dubock P., "Numerical Analysis of Forward and Reverse Biased Potential Distribution in a Two-Dimensional P-N Junction with Applications to Capacitance Calculations", Electron. Lett., Vol.5, pp.236-238, 1969.

/42/ Duff I.S., "A Survey of Sparse Matrix Research", Proc. IEEE, Vol. 65, pp.500-535, 1977.

/43/ Duff I.S., "Practical Comparison of Codes for the Solution of Sparse Linear Systems", A.E.R.E. Harwell, Oxfordshire, 1979.

/44/ Dupont T., Kendall R.D., Rachford H.H., "An Approximate Factorization Procedure for Solving Self-Adjoint Elliptic Difference Equations", SIAM, J. Num. Anal., Vol.5, pp.559-573, 1968.

/45/ Engl W.L., Dirks H., "Numerical Device Simulation Guided by Physical Approaches", Proc. NASECODE I Conf., pp.65-93, 1979.

/46/ Fichtner W., "Untersuchungen an MOS-Transistoren", Dissertation, Technische Universität Wien, 1978.

/47/ Forsythe G.E., Wasow W.R., "Finite Difference Methods for Partial Differential Equations", Wiley, New York, 1960.

/48/ Fossum J.G., "Computer-Aided Numerical Analysis of Silicon Solar Cells", Solid-State Electron., Vol.19, pp.269-277, 1976.

/49/ Fox L., "Finite-Difference Methods in Elliptic Boundary-Value Problems", The State of the Art in Numerical Analysis, Academic Press, London, pp.799-881, 1977.

/50/ Gaensslen F.H., Jaeger R.C., "Temperature Dependant Threshold Behaviour of Depletion Mode MOSFET's", Solid-State Electron., Vol.22, pp.423-430, 1979.

/51/ Gansner M., Ilegems M., Schwob P., Dutoit M., "Modelisation des Structures Microelectroniques de Petites Dimensions", Proc. Journees d'Electronique, pp.93-105, 1980.

/52/ Gaur S.P., "Performance Limitations of Silicon Bipolar Transistors", IEEE Trans. Electron Devices, Vol.ED-26, pp.415-421, 1979.

/53/ Gaur S.P., "Two-Dimensional Carrier Flow in a Transistor under Nonisothermal Conditions", IBM J. Res. Dev., Vol.21, pp.306-314, 1977.

/54/ Gaur S.P., Navon D.H., "Two-Dimensional Carrier Flow in a Transistor Structure under Nonisothermal Conditions", IEEE Trans. Electron Devices, Vol.ED-23, pp.50-57.

/55/ Gibbons J., Johnson W.S., Mylroie S.W., "Projected Range Statistics", Halstead Press, Strandsberg, 1975.

/56/ Greenfield J.A., Dutton R.W., "Nonplanar VLSI Device Analysis Using the Solution of Poisson's Equation", IEEE Trans. Electron Devices, Vol.ED-27, pp.1520-1532, 1980.

/57/ Greenfield J.A., Hansen S.E., Dutton R.W., "Two-Dimensional Analysis for Device Modeling", Stanford Technical Report No. G-201-7, 1980.

/58/ Gummel H.K., "A Self-Consistent Iterative Scheme for One-Dimensional Steady State Transistor Calculations", IEEE Trans. Electron Devices, Vol.ED-11, pp.455-465, 1964.

/59/ Hamilton D.J., Howard W.G., "Basic Integrated Circuit Engineering", McGraw-Hill Kogakusha, Tokyo, 1975.

/60/ Hart J.F., Cheney E.W., Lawson C.L., Maehly H.J., "Computer Approximations", Wiley, New York, 1968.

/61/ Heimeier H.H., "A Two-Dimensional Numerical Analysis of a Silicon N-P-N Transistor", IEEE Trans. Electron Devices, Vol.ED-20, pp.708-714, 1973.

/62/ Heimeier H.H., "Zweidimensionale numerische Lösung eines nichtlinearen Randwertproblems am Beispiel des Transistors in stationaren Zustand", Dissertation, Technische Hochschule Aachen, 1973.

/63/ Holt C.A., "Electronic Ciruits", Wiley, New York, 1978.

/64/ Jacoboni C., Canali C., Ottaviani G., Quaranta A.A., "A Review of Some Charge Transport Properties of Silicon", Solid-State Electron., vol.20, pp.77-89, 1977.

/65/ Jaggi R., Weibel H., "High-Field Electron Drift Velocities and Current Densities in Silicon", Helv.Phys.Acta., Vol.42, pp.631-632, 1969.

/66/ Jesshope C.R., "Bipolar Transistor Modelling with Numerical Solutions to the 2-Dimensional DC and Transient Problems", Dissertation, University of Southampton, 1976.

/67/ Jesshope C.R., Zaluska E.J., Kemhadjian H.A., "Comparison of Numerical-Solution Methods for 2-Dimensional Bipolar-Transistor-Analysis Algorithm", Electron. Lett., Vol.11, pp.285-286, 1975.

/68/ Kahng D., Atalla M.M., "Silicon-Silicon Dioxide Field Induced Surface Devices", IRE-AIEE, Solid-State Device Res. Conf., 1960.

/69/ Kani K., "A Survey of Semiconductor Device Analysis in Japan", Proc. NASECODE I Conf., pp.104-119, 1979.

/70/ Kennedy D.P., O'Brien R.R., "Analysis of the Impurity Atom Distribution Near the Diffusion Mask for a Planar p-n Junction", IBM J. Res. Dev., Vol.9, pp.179-186, 1965.

/71/ Kennedy D.P., Murley P.C., "Steady State Mathematical Theory for the Insulated Gate Field Effect Transistor", IBM J. Res. Dev., Vol.17, pp.2-12, 1973.

/72/ Kennedy D.P., O'Brien R.R., "Two-Dimensional Mathematical Analysis of a Planar Type Junction Field-Effect Transistor", IBM J. Res. Dev., Vol.13, pp.662-674, 1969.

/73/ Knuth D.E., "Fundamental Algorithms", Addison-Wesley, Reading, 1977.

/74/ Knuth D.E., "Seminumerical Algorithms", Addison-Wesley, Reading, 1969.

/75/ Kotani N., Kawazu S., "Computer Analysis of Punch-Through in MOSFET's", Solid-State Electron., Vol. 22, pp.63-70, 1979.

/76/ Lavine J.P., Burkey B.C., "Extensions of the Scharfetter-Gummel Approach to Charge Transfer", Solid State Electron., Vol.23, pp.75-77, 1980.

/77/ Lawson C.L., Handson R.J., Kincaid D.R., Krogh F.T., "Basic Linear Algebra Subprograms for FORTRAN Usage", Sandia Laboratories Rep., No.SAND77-0898, 1977.

/78/ Lee H., Sansbury J.D., Dutton R.W., Moll J.L., "Modeling and Measurement of Surface Impurity Profiles of Laterally Diffused Regions", IEEE J. Solid-State Circuits, Vol.SC-13, pp.455-461, 1978.

/79/ Lee H., "Two-Dimensional Impurity Diffusion Studies: Process Models and Test Structures for Low-Concentration Boron Diffusion", Stanford Technical Report No.G201-8, 1980.

/80/ Li S.S., Thurber W.R., "The Dopant Density and Temperature Dependence of Electron Mobility and Resistivity in n-Type Silicon", Solid-State Electron., Vol.20, pp.609-616, 1977.

/81/ Liu S., Hoefflinger B., Pederson D.O., "Interactive Two-Dimensional Design of Barrier Controlled MOS Transistors", IEEE Trans. Electron Devices, Vol.ED-27, pp.1550-1558, 1980.

/82/ Loeb H.W., Andrew R., Love W., "Application of 2-Dimensional Solutions of the Shockley-Poisson Equation to Inversion-Layer M.O.S.T. Devices", Electron. Lett., Vol.4, pp.352-354, 1968.

/83/ Manck O., Heimeier H.H., Engl W.L., "High Injection in a Two-Dimensional Transistor", IEEE Trans. Electron Devices, Vol.ED-21, pp.403-409, 1974.

/84/ Manck O., "Numerische Analyse des Schaltverhaltens eines zweidimensionalen bipolaren Transistors", Dissertation, Technische Hochschule Aachen, 1975.

/85/ Manck O., Engl W.L., "Two-Dimensional Computer Simulation for Switching a Bipolar Transistor out of Saturation", IEEE Trans. Electron Devices Vol.ED-24, pp.339-347, 1975.

/86/ Marhsak A.H., Shrivastava R., "Law of the Junction for Degenerate Material with Position-Dependent Band Gap and Electron Affinity", Solid-State Electron., Vol.22 pp., 567-571, 1979.

/87/ Marshal D., "Die Numerische Lösung partieller Differential-gleichungen", Bibliographisches Institut, Mannheinm, 1976.

/88/ Masuda H., Nakai M., Kubo M., "Characteristics and Limitation of Scaled-Down MOSFET's due to Two-Dimensional Field-Effect", IEEE Trans. Electron Devices, Vol.ED-26, pp.980-986, 1979.

/89/ Mertens R.P., Van Meerbergen J.L., Nijs J.F., Van Overstraeten R.J., "Measurement of the Minority-Carrier Transport Parameters in Heavily Doped Silicon", IEEE Trans. Electron Devices, Vol.ED-27, pp.949-955, 1980.

/90/ Mock M.S., "A Two-Dimensional Mathematical Model of the Insulated-Gate Field-Effect Transistor", Solid-State Electron., Vol.16, pp.601-609, 1973.

/91/ Mock M.S., "The Charge-Neutral Approximation and Time Dependent Simulation", Proc. NASECODE I Conf., pp.120-135, 1979.

/92/ Mock M.S., "Transport Equations in Heavily Doped Silicon, and the Current Gain of a Bipolar Transistor", Solid-State Electron., Vol. 16, pp.1251-1259, 1973.

/93/ Nakagawa A., "One-Dimensional Device Model of the npn Bipolar Transistor Including Heavy Doping Effects under Fermi Statistics", Solid-State Electron., Vol. 22, pp.943-949, 1979.

/94/ Nussbaum A., "Inconsistencies in the Original Form of the Fletcher Boundary Conditions", Solid-State Electron., Vol.21, pp.1178-1179, 1978.

/95/ Ochi S., Okabe T., Yoshida I., Yamaguchi K., Nagata K., "Computer Analysis of Breakdown Mechanism in Planar MOSFET's", IEEE Trans. Electron Devices, Vol.ED-27, pp.,399-400, 1980.

/96/ OH S.Y., Ward D.E., Dutton R.W., "Transient Analysis of MOS Transistors", IEEE Trans. Electron Devices, Vol.ED-27, pp.1571-1578, 1980.

/97/ Oka H., Nishiuchi K., Nakamura T., Ishikawa H., "Computer Analysis of a Short-Channel BC MOSFET", IEEE Trans. Electron Devices, Vol.ED-27, pp.1514-1520, 1980.,

/98/ Oka H., Nishiuchi K., Nakamura T., Ishikawa H., "Two-Dimensional Numerical Analysis of Normally-Off Type Buried Channel MOSFET's", Proc. International Electron Devices Meeting, pp.30-33, 1979.

/99/ Ortega J.M., Rheinboldt W.C., "Iterative Solution of Nonlinear Equations in Several Variables", Academic Press, New York, 1970.

/100/ Polsky B.S., Rimshans J.S., "Comparison of Different Methods for Numerical Simulation of Transient Processes in Bipolar Semiconductor Devices", Solid-State Electron., Vol. 23, pp.183-185, 1980.

/101/ Reiser M., "A Two-Dimensional Numerical FET Model for DC, AC, and Large-Signal Analysis", IEEE Trans. Electron Devices, Vol.ED-20, pp.35-44, 1973.

/102/ Reiser M., "Difference Methods for the Solution of the Time-Dependent Semiconductor Flow-Equations", Electron. Lett., Vol.7, pp.353-355, 1971.

/103/ Reiser M., "Two-Dimensional Analysis of Substrate Effects in Junction F.E.T.s", Electron. Lett., Vol.6, pp. 493-494, 1970.

/104/ Rheinboldt W.D., "Methods for Solving Systems of Nonlinear Equations", SIAM, Philadelphia, 1974.

/105/ Rokus A., "Zur numerischen Losung linearer pentagonaler Gleichungssysteme hohen Ranges", Diplomarbeit, Technische Universität Wien, 1980.

/106/ Runge H., "Distribution of Implanted Ions under Arbitrarily Shaped Mask Edges", Phys. Status Solidi, Vol.(a)39, pp.595-599, 1977.

/107/ Ryssel H., Hamberger K., Hoffmann K., Prinke G., Dümcke R., Sachs A., "Simulation of Doping Processes", IEEE Trans. Electron Devices, Vol.ED-27, pp.1484-1492, 1980.

/108/ Ryssel H., Ruge I., "Ionenimplantation", Teubner, Stuttgart, 1978.

/109/ Sabnis A.G., Clemens J.T., "Characterization of the Electron Mobility in the inverted (100) SI-Surface", Proc. International Electron Devices Meeting, pp.18-21, 1979.

/110/ Scharfetter D.L., Gummel H.K., "Large-Signal Analysis of a Silicon Read Diode Oscillator", IEEE Trans. Electron Devices, Vol.ED-16, pp.64-77, 1969.

/111/ Seeger K., "Semiconductor Physics", Springer, Wien, 1973.

/112/ Selberherr S., Schültz A., Pötzl H., "MINIMOS - Zweidimensionale Modellierung von MOS-Transistoren", Elektronikschau, Vol.9, pp.18-23, Vol.10, pp.54-58, 1980.

/113/ Selberherr S., Fichtner W., Pötzl H.W., "MINIMOS - a Program Package to Facilitate MOS Device Design and Analysis", Proc. NASECODE I Conf., pp.275-279, 1979.

/114/ Selberherr S., Schültz A., Pötzl H.W., "MINIMOS - a Two-Dimensional MOS Transistor Analyzer", IEEE Trans. Electron Devices, Vol.ED-27, pp.1540-1550, 1980.

/115/ Selberherr S., Guerrero E., "Simple and Accurate Representation of Implantation Parameters by Low Order Polynomals", Solid-State Electron., Vol.24, pp.xxx-xxx, 1981.

/116/ Siebert H., "Höhere Fortran Programmierung", De Gruyter, Berlin, 1974.

/117/ Slotboom J.W., "Computer-Aided Two-Dimensional Analysis of Bipolar Transistors", IEE Trans. Electron Devices, Vol.ED-20, pp.669-679, 1973.

/118/ Slotboom, J.W., "Iterative Scheme for 1- and 2-Dimensional D.C.-Transistor Simulation", Electron. Lett., Vol.5, pp.677-678, 1969.

/119/ Slotboom J.W., De Graaff H.C., "Measurements of Bandgap Narrowing in Si Bipolar Transistors", Solid-State Electron., Vol.19, pp.857-862, 1976.

/120/ Smith G.D., "Numerical Solution of Partial Differential Equations: Finite Difference Methods", Clarendon Press, Oxford, 1978.

/121/ Smith R.A., "Semiconductors", Cambridge University Press, Cambridge, 1978.

/122/ Stoer J., Bulirsch R., "Einführung in die Numerische Mathematik II", Springer, Berlin, 1978.

/123/ Stone H.L., "Iterative Solution of Implicit Approximations of Multidimensional Partial Differential Equations", SIAM J.Numer.Anal., Vol.5, pp.530-558, 1968.

/124/ Sun S.C., Plummer J.D., "Electron Mobility in Inversion and Accumulation Layers on Thermally Oxidized Silicon Surfaces", IEEE Trans. Electron Devices, Vol.ED-27, pp.1497-1508, 1980.

/125/ Sutherland A.D., "A Two-Dimensional Computer Model for the Steady-State Operation of MOSFET's", US. Army Electronics Command Res. and Dev. Techn. Rep., ECOM-75-1344-F, 1977.

/126/ Sutherland A.D., "An Algorithm for Treating Interface Surface Charge in the Two-Dimensional Discretization of Poisson's Equation for the Numerical Analysis of Semiconductor Devices such as MOSFET's", Solid-State Electron., Vol. 23, pp.1085-1087 1980.

/127/ Sutherland A.D., "On the Use of Overrelaxation in Conjunction with Gummel's Algorithm to Speed the Convergence in a Two-dimensional Computer Model for MOSFET's", IEEE Trans. Electron Devices, Vol.ED-27, pp.1297-1298, 1980.

/128/ Sze S.M., "Physics of Semiconductor Devices", Wiley, New York, 1969.

/129/ Takacs D., Schwabe U., Bürker U., "The Influence of Temperature on the Tolerances of MOS-Transistors in a One-Micrometer Technology", Proc. International Electron Devices Meeting, pp.569-573, 1980.

/130/ Tielert R., "Two-Dimensional Numerical Simulation of Impurity Redistribution in VLSI Processes", IEEE Trans. Electron Devices, Vol.ED-27, pp.1479-1483, 1980.

/131/ Toyabe T., Yamaguchi K., Asai S., Mock M., "A Numerical Model of Avalanche Breakdown in MOSFET's", IEEE Trans. Electron Devices, Vol.ED-25, pp.275-279, 1978.

/132/ Toyabe T., Asai S., "Analytical Model of Threshold Voltage and Breakdown Voltage of Short-Channel MOSFET's derived from Two-Dimensional Analysis", IEEE Trans. Electron Devices, Vol.ED-26, pp.453-461, 1979.

/133/ Troutman R.R., "VLSI Limitations from Drain-Induced Barrier Lowering.", IEEE Trans. Electron Devices, Vol.ED-26, pp.461-469, 1979.

/134/ Van Roosbroeck W.V., "Theory of Flow of Electrons and Holes in Germanium and Other Semiconductors", Bell Syst. Techn. J., Vol.29, pp.560-607, 1950.

/135/ Van Vliet K.M., "On Fletcher's Boundary Conditions", Solid-State Electron., Vol.22, pp.443-444, 1979.

/136/ Van Vliet K.M., "The Shockley-Like Equations for the Carrier Densities and the Current Flows in Materials with a Nonuniform Composition", Solid-State Electron., Vol.23, pp.49-53, 1980.

/137/ Vandorpe D., Borel J., Merckel G., Saintot P. "An Accurate Two-Dimensional Numerical Analysis of the MOS Transistor", Solid-State Electron., Vol.15, pp.547-557, 1972.

/138/ Vandorpe D., Xuong N.H., "Mathematical 2-Dimensional Model of Semiconductor Devices", Electron. Lett., Vol.7, pp.47-50, 1971.

/139/ Varga R.S., "Matrix Iterative Analysis", Prentice-Hall, Englewood Cliffs, 1962.

/140/ Wachspress E.L., "Iterative Solution of Elliptic Systems", Prentice-Hall, Englewood Cliffs, 1966.

/141/ Warner D.D., Wilson C.L., "Two-Dimensional Concentration Dependent Diffusion", Bell Syst. Techn. J., Vol.59, pp.1-41, 1980.

/142/ Yamaguchi K., "Field-Dependant Mobility Model for Two-Dimensional Numerical Analysis of MOSFET's", IEEE Trans. Electron Devices, Vol.ED-26, pp.1068-1074, 1979.

/143/ Yamaguchi K., Toyabe T., Kodera H., "Two-Dimensional Analysis of Triode-Like Operation of Junction Gate FET's", IEEE Trans. Electron Devices, Vol.ED-22, pp.1047-1049, 1975.

/144/ Yokoyama K.Y., Yoshii A., Horiguchi S., "Threshold Sensitivity Minimization of Short-Channel MOSFET's by Computer Simulation", IEEE Trans. Electron Devices, Vol.ED-27, pp.1509-1514, 1980.

/145/ Yoshii A., Horiguchi S., Sudo T., "A Numerical Analysis for Very Small Semiconductor Devices", Proc. International Solid-State Circuits Conf., pp.80-81, 1980.

/146/ Yourdon E., Constantine L.L., "Structured Design", Prentice-Hall, Englewood Cliffs, 1979.

/147/ Zaluska E.J., Dubock P.A., Kemhadhan H.A., "Practical 2-Dimensional Bipolar-Transistor-Analysis Algorithm", Electron. Lett., Vol.9., pp.599-600, 1973.

APPENDIX C THE STRUCTURE OF MINIMOS

The structure of the computer program MINIMOS is discussed from the following point of view; that a segmentation of the pertinent object code, which is calculated to use the smallest amount of main memory, can be simply accomplished.

Figure C-1 shows a diagram for such a segmentation with symbolic segment names. The table in figure C-2 further shows which subroutines belong to each segment in the diagram of figure C-1.

The structure of the program uses a four level segmentation with a very small root segment (S00) resident in main memory. This root segment includes only the main program, the initialization of the physical and mathematical constants and a small subprogram which is necessary for all lower order segments. The main program is realized as a driver program, that means there are no essential calculations performed, instead it has only the specific task of calling the other subprograms (segments). This structure was implemented as a consequence of future programming goals, in order to obtain the highest possible level of modularity and thereby future programming changes such as enhancements can be easily carried out /146/.

The first sublevel contains two segments, which are logically completely independent. The segment S01 converts and analyses the input directives, which are explained in detail in Appendix A. The support subroutine in this segment is SUB01; it represents the framework of a directive compiler. The segment S02 is the driver segment for all mathematical calculations. The support subroutine here is SUB02.

All of the essential calculations and the output of the results follow in the next level of segmentation. In segment S03, with SUB02A as the support subroutine, all of the input data is normalized (TEST), an initial grid is calculated (GRID), the doping profile is calculated (CONCNT), and the intitial solution is calculated (INIT).

In segment S04-SUB02B is the relevant subroutine - the initial solution and the grid are refined as much as possible in order that the actual two

dimensional solution, which is very time intensive, can be obtained. The procedure is as follows: First all of the grid dependent quantities are calculated (POICO), the boundary conditions are imposed (BOUND), then the two dimensional Poisson's equation and a one dimensional continuity equation (SIMUL) are solved and the grid along with the resulting solution are checked for optimality (CHECK). When the grid satisfies certain conditions which are deduced from the solution, then this segment is released. If the grid is not satisfactory, it is regenerated. Grid lines will be added, moved or removed. This procedure is called adaptive grid generation /12/, /74/. Consequently the last solution will be carried over (SAVE), the doping proifile for the new grid is calculated (CONCNT) and furthermore at the beginning of the segment, the calculations are begun for the new grid dependent coefficients. This procedure is repeated until the grid satisfies all necessary conditions.

The essential two dimensional calculations follow in segment S05. The support, and realistic driver subroutine here is SUB02C. First the mobility distribution is calculated (MOB) and the last obtained solution is tested for physical and mathematical convergence (CONV). This segment is released when convergence is reached, otherwise the two dimensional current continuity equation for minority carriers will be solved (CEQMIN), similarly for Poisson's equation (POIEQ), and then the program returns to the beginning of the segment, until a mathematically consistent solution is found.

The last segment in this level, designated as S06, prints all results on the line printer. The responsible subroutine is named SUB02D.

If more operating points were to be calculated with MINIMOS, the execution would return to segment S04.

The segmentation shown in the fourth level of figure C-1 is not absolutely necessary. In many cases it cannot be simply realized; it is only given in order to make the fine structure simple where needed and possible. The given structuring was tested on a CDC computer with the help of the system programs SEGLOAD /26/ with excellent results. The necessary memory was reduced by about 25 percent by the segmentation.

At this point it should also be mentioned, that all subroutines in figure C-2, which begin with ZZZ and are in bold type, are system dependent. One must modify these when MINIMOS is installed in different computers. In general, this presents no overall problem, in that these routines are very short. A detailed description of the necessary adaptations can be taken from the comments in the source text of the programs.

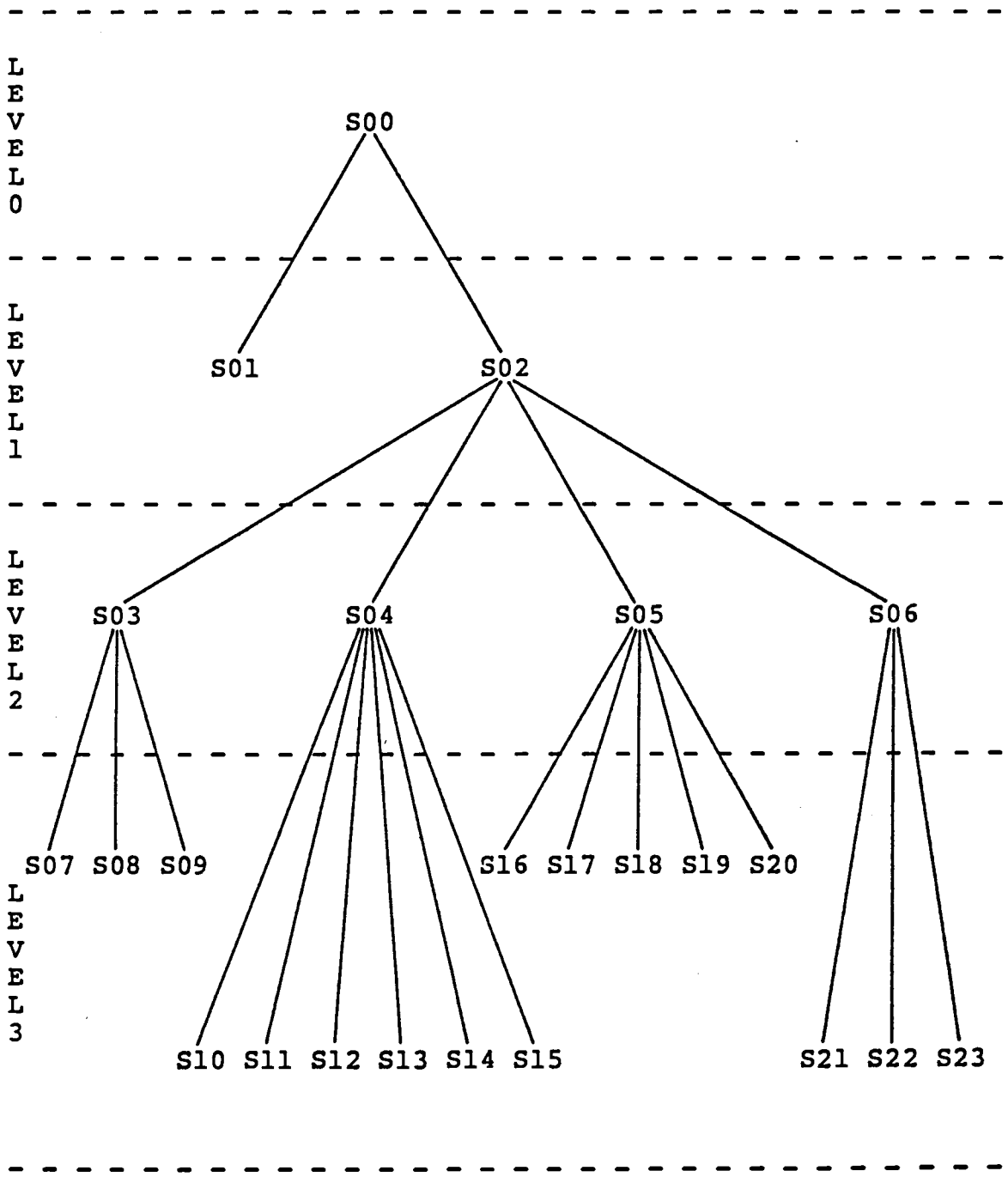


Figure C-1: The segmentation diagram of MINIMOS

S00	S01	S02	S03	S04	S05	S06	S07
(main)	SUB01	SUB02	SUB02A	SUB02B	SUB02C	SUB02D	TEST
blkdat	PARSE	REFERI	CONCNT		QUAPHI	QUAPHI	
WRTHED	REALVA	SETUP	IMPPAR		CONV	CHARGE	
REFERZ	LOGIVA	MEMORY	IMPSTA			PRARAY	
REFERT	IFEQ	ERFXX	CUBF				
ABTMOS	ERRORA	REFERR					
ZZZSEC	ERRORB	SETMEM					
	ERRORC	ZZZOPE					
	ERRORD	ZZZPUT					
	ERRMSG	ZZZGET					
	CLASH	ZZZCLO					
	ZZZDAT						

S08	S09	S10	S11	S12	S13	S14	S15
GRID	INIT	POICO	BOUND	SIMUL	SAVE	CHECK	CONCNT
				POISIP			IMPPAR
							IMPSTA
							CUBF

S16	S17	S18	S19	S20	S21	S22	S23
MOB	CEQMIN	POIEQ	CURL	BINAER	FIELDL	FIELDT	CURT
	CEQSIP	POISIP					
	MOVMEM						

Figure C-2: The segments and their contents

B I O G R A P H Y

I was born August 3, 1955 the son of Johannes and Josefine Selberherr in Klosterneuburg, Austria. After attending the Primary School in Abstetten, Austria I entered the modern language High School in Tulln, Austria from which I graduated on June 7, 1973.

In the winter semester of the 1973/74 school year, I began the study of Industrial Electronics and Control Theory at the Technical University of Vienna. On June 21, 1978 I completed these studies with the title of Diplom-Engineer.

From September 4, 1978 to May 30, 1980 I was a research assistant in the Institute for Physical Electronics. Since June 1, 1980 I have been a University assistant at the same institution (presently the Institute for General Electrical Engineering and Electronics).