EMPIRICAL MOSFET MODELS FOR CIRCUIT SIMULATION

by

J. L. Burns

Memorandum No. UCB/ERL M84/43

25 May 1984

(cover)

Empirical MOSFET Models for Circuit Simulation

Jeffrey L. Burns

Department of Electrical Engineering and Computer Sciences
Electronics Research Laboratory
University of California, Berkeley, CA

EMPIRICAL MOSFET MODELS FOR CIRCUIT SIMULATION

by

J. L. Burns

Memorandum No. UCB/ERL M84/43

25 May 1984

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

ABSTRACT

Empirical models for MOS transistors have been investigated in this project. The work has resulted in two models, one based upon a 2-dimensional table and several 1-dimensional functions, and one based upon 1-dimensional functions only. The dimensionality refers to the number of independent variables present.

The 1-dimensional functions are cubic spline fitting functions, which are continuous and differentiable. Interpolation techniques which are computationally efficient and which have physical significance are used for the 2-dimensional table. The interpolating functions and/or the spline functions are used to compute well-behaved partial derivatives of the MOSFET drain current with respect to its node-pair voltages. Proper behavior of the partial derivatives is necessary to insure simulator convergence.

The empirical models have been installed in the SPICE2 circuit simulation program. The models are 2-4 times faster to evaluate than analytical models of comparable accuracy, with very low requisite storage compared to other empirical modeling schemes.

ACKNOWLEDGEMENTS

The author would like to thank his advisors, Professor D.O. Pederson and Professor A.R. Newton, for their support and guidance throughout the course of this project. The author especially values being granted the freedom to pursue this research in the directions of his choosing.

Numerous discussions with A. Vladimirescu are greatly appreciated, as are the help and encouragement of the author's fellow students in the Berkeley CAD group.

The author is grateful to Professor J. Choma and Professor N.C. Luhmann for interesting him in continuing his education at Berkeley.

The author thanks his friends and family for their support and encouragement, and Kris and Nicky for providing a very pleasant distraction.

The financial support of the MICRO project of the University of California, Linkabit Corp., SeeQ Technology Inc., and the Semiconductor Research Corporation are acknowledged.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The majority of integrated circuits (ICs) are composed entirely, or nearly so, of transistors. The utility of computer simulation as a tool for aiding in the design of ICs therefore depends on the models used to represent the transistors. In particular, the accuracy and computational efficiency of the models directly affect the corresponding accuracy and speed of the simulation.

This report presents two *empirical* metal-oxide-semiconductor field-effect transistor (MOSFET) models for use in circuit simulation programs. The models are *semi-physical* models, meaning that they are constructed in a manner that exploits the physics of the MOS device.

Like all circuit simulator models, empirical models are passed a set of voltages from the simulation program which specifies the operating point of the device. From these voltage inputs, the model subroutine returns the element values for an appropriate incremental MOSFET circuit model. Some empirical models store the necessary information in tables, and are called table look-up models. Other empirical models consist of numerical functions that are used to curve-fit the data, and are called function-fit models. Two empirical models are described in this report. One develops the element values from a data set which is stored partially in a table and partially in fitting functions, whereas the other is a function-fit model.

Most semiconductor device models comprise one or more nonlinear equations derived from physical principles, and are called *analytical* models.

A second class of device models consists of models which analytically represent the first-order behavior, but account for higher-order effects through the introduction of empirical parameters. These models are usually referred to as *semi-empirical* models, an example of which is the SPICE2 LEVEL-3 MOSFET model [1]. Modern MOSFETs display the effects of many complicated physical phenomena. Thus, accurate analytical and semi-empirical models typically consist of several complicated nonlinear equations characterized by many physical parameters or empirical parameters or both. Finding values for the parameters is often a difficult problem. The physical parameters frequently must be used in a curve-fitting fashion because of the approximations used in the derivation of the model equations. Also, each parameter value must contain a large amount of compressed information.

There are other problems associated with analytical and semi-empirical models, although the problems are potentially less severe for semi-empirical models.[1] For example, these models must be regularly revised to account for process changes and changes in the technology. Because the evolution of a technology precedes its understanding in device-physics terms, the resulting analytical models are seldom optimal in their fitting ability. Also, good-quality analytical models are often expensive to evaluate.

Empirical models have several potential advantages. The data set can be made very general and need not contain any process or technology-dependent parameters. As a result, empirical models do not need revision as the process and/or technology evolves. The data set is large relative to an analytical model's parameter set. This implies that the data set is easier to

[1]For the remainder of this chapter the term analytical refers to semi-empirical as well.

determine than a parameter set, due to the lesser degree of information compression. The large data set also makes the overall fit of an empirical model insensitive to local errors in the data set. An error in a single parameter value of an analytical model can strongly degrade the overall fit. Empirical models can be made arbitrarily accurate via increasing the size of the data set. Empirical models are often faster to evaluate than equivalent analytical models.

The two empirical models which appear in Chapters 4 and 5 of this report differ from one another in terms of storage dimensionality. The data set of the two-dimensional (2-d) empirical model is partially contained in a table whose entries are referenced by two independent node-pair voltages, with the remainder contained in functions referenced by one independent node-pair voltage. The one-dimensional (1-d) empirical model's data set is stored in a collection of functions that depend only on a single node-pair voltage. The significance of the storage dimensionality can be appreciated by noting that each element in a MOSFET circuit model is generally a function of *three* independent node-pair voltages. A quantity specified by three independent variables needs a storage allocation proportional to n^3 for n points per dimension. Quantities which depend on one or two independent variables require storage allocations proportional to n or n^2, respectively.

The empirical models of this research possess an advantage in computational efficiency over accurate analytical models, by a factor of about two for the 2-d model, and by a factor of about four for the 1-d model. The ease of fitting these two empirical models to $I-V$ data is demonstrated also.

This report is organized as follows: in Chapter 2, the problem of non-linear device modeling for circuit simulation is addressed, and a description of the MOS transistor is given. The basics of MOSFET modeling, via analytical models and via empirical models, is covered in Chapter 3. The two empirical models which are the subject of this research are presented in Chapters 4 and 5. The report concludes with Chapter 6, containing results.

CHAPTER 2

MODELING NONLINEAR DEVICES FOR CIRCUIT SIMULATION

2.1. INTRODUCTION

Circuits containing nonlinear elements, such as MOSFETs, are described by systems of nonlinear algebraic equations for DC analysis, and by systems of nonlinear differential equations for transient analysis. Circuit simulation programs approximate the solution of the nonlinear equations by a *succession* of solutions of a *linearized* system of equations generated from the original nonlinear system. The nonlinear devices are modeled in the associated linear equation representation of the circuit by incremental linear models called *companion models* [2]. The values of the elements which comprise the MOSFET companion model are generated by the empirical models described in this report.

In this chapter, the usual technique used in the linearization of the circuit equations is presented. The physical behavior of MOSFETs is described, illustrating the nonlinear behavior these elements display. The MOSFET companion model is then derived.

2.2. NONLINEAR DEVICES AND SIMULATION

2.2.1. Methodology of Simulation Circuit simulation programs perform the time-domain transient analysis of electronic circuits by the sequence of steps illustrated in Figure 2.1 [3]. For this research, the segment of Fig. 2.1 which is important is the box labeled "Linearize Semiconductor Devices

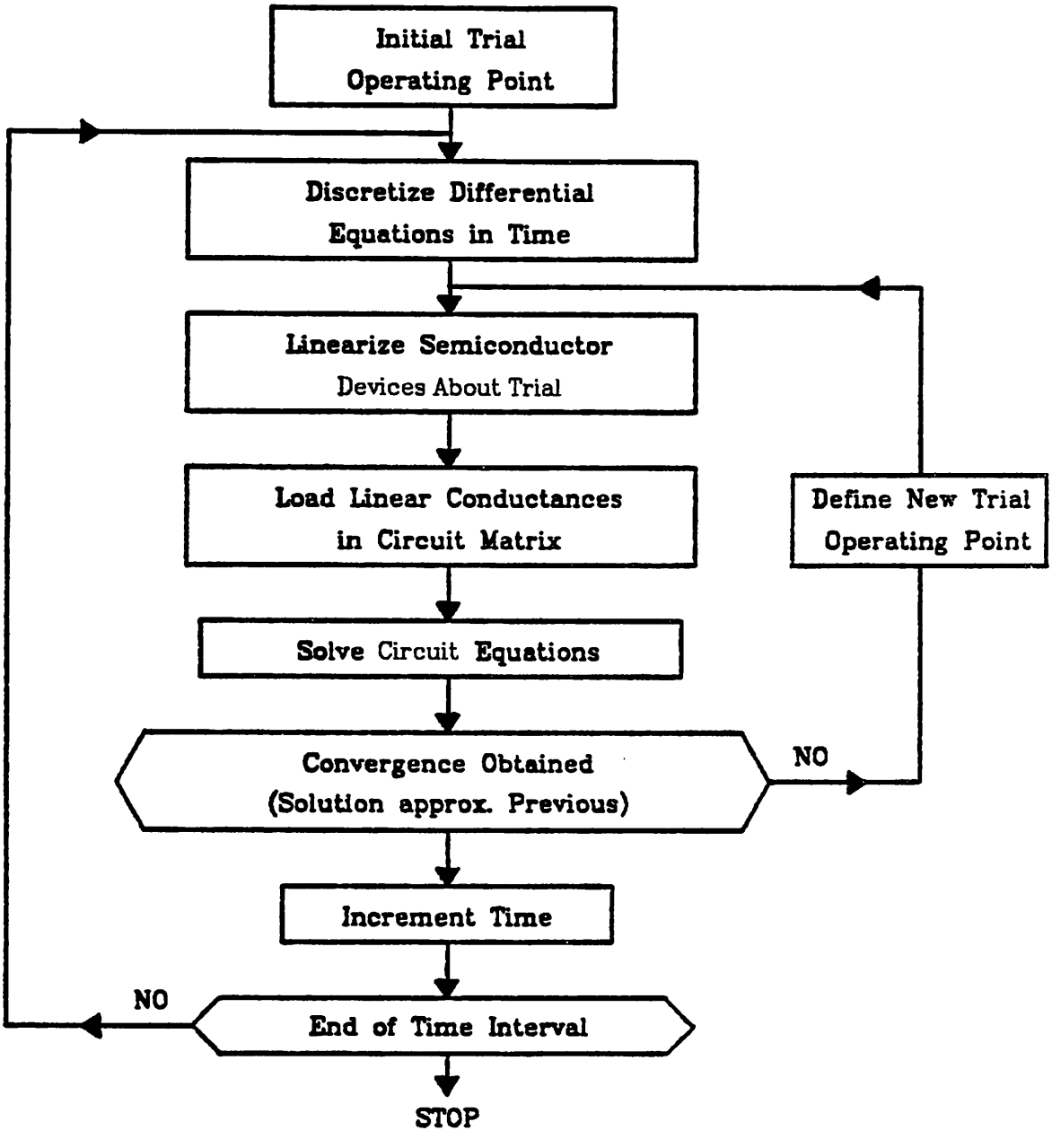

Figure 2.1 Flowchart for Transient Analysis

About Trial". The linearization corresponds to developing a linear incremental circuit model for each nonlinear element at its respective operating point.

A significant consideration in model development is insuring that the overall properties of the circuit simulation program are not adversely affected by the new model. In particular, a model is not acceptable if it does not preserve the convergence properties of the simulation program. The specific requirements for convergence of the linearization algorithm used in most circuit simulation programs are described later in this chapter. Testing to date shows that the two empirical models presented in this report meet the requirement that the convergence properties of the simulator are not degraded.

2.2.2. Performance of Simulators The time required for a circuit simulation is significant, especially for large circuits. Practical cases include circuits of up to several thousand nonlinear devices, which require cpu times on the order of hours or even days. The role of modeling nonlinear elements with respect to the time needed to perform a simulation is thus important.

The cpu time spent on a circuit simulation can be broken into two components: the per-iteration time required for one device model evaluation (t_d), and the time required for one linear equation solution (t_s). One way of writing the total time in terms of these two characteristic times is [4]

$$T = n_i(n_d t_d + n_s t_s) + \text{overhead},$$

where n_i, n_d, and n_s are the number of iterations, devices, and equations, respectively. In this context, t_d is the term of concern.

The fraction of T taken by model evaluation varies with the simulator, the computer, and the size of the circuit. In small to medium-sized circuits

where accurate analytical or semi-empirical MOSFET models are used, well over half of T is consumed by evaluation of the model equations. For example, a low-pass filter circuit with 70 MOSFETs has been analyzed on the Cray 1 computer with the circuit simulation program SPICE2 [4]. The model used for the transistors is the LEVEL-3 [5] semi-empirical model built into the program. The LEVEL-3 model is the more efficient of the two accurate models in SPICE2 [1]. The time needed for model evaluation accounts for over 73% of T in this example [4].

A goal of this project has been to develop MOSFET models which have improved t_d over equivalent analytical or semi-empirical models. This goal has been achieved; the 2-d model is about two times faster to evaluate than a good analytical model, whereas the 1-d model is about four times faster.

2.2.3. Linearization of Circuit Equations The Newton-Raphson (NR) algorithm is the method used in virtually all circuit simulation programs to perform the linearization of the nonlinear circuit equations. The NR method is used because of its advantage in rate of convergence over other linearization methods, and because the factors which tend to lead to nonconvergence of the NR algorithm can be eliminated without any disadvantages as compared to other methods [3].

The NR algorithm is actually the result of approximating a nonlinear function by a truncated Taylor series [6]. In one variable the problem to be solved is

$$f(x) = 0, \tag{2.1}$$

where $f(x)$ is a nonlinear function and x is a variable. At the solution the value for x is a root of Eq.(2.1). For an x close to the solution, say x_0, the Taylor series expansion about x_0 is

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x-x_0)^2}{2!}f''(x_0) + \cdots . \tag{2.2}$$

Equation (2.2) can be truncated above the linear term and substituted into Eq.(2.1) to yield

$$0 = f(x_0) + (x^* - x_0)f'(x_0),$$

which can be rewritten as

$$x^* = x_0 - \frac{f(x_0)}{f'(x_0)} . \tag{2.3}$$

Here, x^* is an approximation to the solution of Eq.(2.1). If x_0 is identified as the j^{th} estimate of the solution, Eq.(2.3) becomes

$$x_{j+1} = x_j - \frac{f(x_j)}{f'(x_j)} . \tag{2.4}$$

Equation (2.4) is Newton's method for a function of only one variable. Convergence of Eq.(2.4) is obtained when x_{j+1} agrees with x_j within a specified tolerance. For n equations in n unknowns, Eq.(2.4) is readily generalized to form the Newton-Raphson method [7]:

$$\bar{x}_{j+1} = \bar{x}_j - \frac{\bar{f}(\bar{x}_j)}{\bar{J}(\bar{x}_j)} , \tag{2.5}$$

where $\bar{x}_j$ is a vector of n variables, and $\bar{f}$ is a vector of n equations. In Eq.(2.5), $\bar{J}$ is the Jacobian matrix, defined as

$$\bar{J}(\bar{x}_j) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_1} \\ \vdots & \cdots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}_j \tag{2.6}$$

The NR algorithm, viz., a truncated Taylor series approximation, can be viewed as a *rule* for constructing a linearized incremental circuit model for a nonlinear element. Such a model is called a *companion model*, as noted previously. Later in this chapter the results of this section are used to

formulate the companion model of a MOS transistor.

2.2.4. Convergence of Newton-Raphson The conditions under which the NR algorithm is guaranteed to converge to a solution are proven in many references, one being [7]. Sufficient conditions for convergence are as follows. If

$$\left| \frac{f(x)f''(x)}{[f'(x)]^2} \right| < 1$$

in an interval about a solution, then the NR algorithm will converge for any initial value of x in the interval. Also required for sufficiency are $f(x)$ and $f'(x)$ continuous in the interval, and f' nonzero there. These are not necessary conditions; the NR method may converge if the conditions are not met.

The key point for simulation is guaranteeing that the nonlinear circuit equations are continuous with continuous first derivatives. If this condition is not met by a device model, the model is usually not acceptable. The equations and derivatives need not be continuous in the strict mathematical sense, however. Due to the error tolerances allowed for in determining when $\bar{x}_{j+1}$ is sufficiently close to (i.e., converged to) $\bar{x}_j$, some discontinuity can be tolerated. The degree of discontinuity must be small enough that the resulting error can be absorbed by the simulation program's error tolerances.

2.3. MOSFET PHYSICS OF OPERATION

In this section, a brief description of MOS transistors is given. The basic structure of the device is outlined, and the results of a simple first-order derivation of its electrical behavior are presented.

2.3.1. Structure and Operation Figure 2.2 depicts the symbol for and cross-sectional view of a MOSFET. The device consists of a metal-oxide-semiconductor (MOS) capacitor that defines a channel region of length L and width W, and two p-n junctions, one at each end of the channel. The transistor in Figure 2.2 is an n-channel MOSFET of the normally off, or

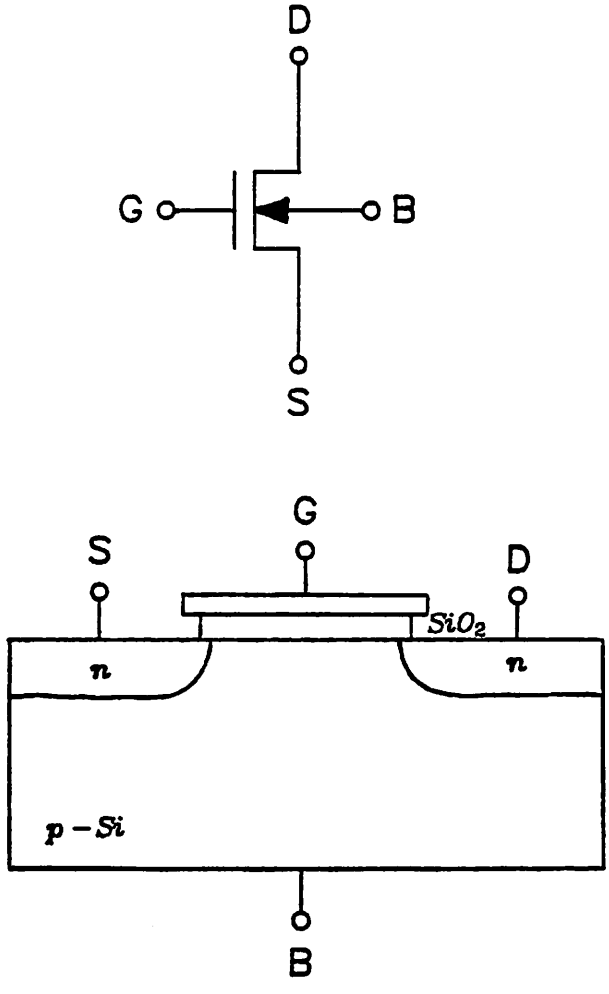

Figure 2.2 Symbol and Cross-Section of a MOSFET.

enhancement-mode type. The n-channel designation refers to the polarity of the carriers present when the transistor is conducting. In an n-channel device, electrons are the carriers; in a p-channel device, holes are the carriers. In an enhancement-mode device, no channel is present in the absence of applied bias at the gate electrode, and the source and drain junctions are thus disconnected. If a sufficiently large positive voltage is applied to the gate, a number of electrons large enough to invert the polarity of the substrate material from p-type to n-type is attracted to the silicon-silicon dioxide interface and the source and drain are electrically connected. Adding bias to the gate thus *enhances* the concentration of carriers in the channel and increases its conductivity. Changing the potential at any of the other non-reference terminals, i.e., the drain or bulk, also changes the concentration of electrons in the channel and modulates the current which flows between the source and the drain. When a conducting channel is present, the MOSFET is said to be operating in *inversion*. In inversion, electron current flows from the drain to the source if the drain voltage is more positive than the source voltage. The current increases sharply with drain-source voltage V_{DS} when V_{DS} is small. A sufficiently large V_{DS} causes the drain current to level off, or *saturate*, at some value. Further increases in V_{DS} do not increase the drain current markedly.[1]

Many other MOSFET types are used, varying mainly in the channel conductivity at zero gate bias. Another common one is the *depletion-mode* transistor, which has a strongly conducting channel for zero gate bias. It is beyond the scope of this report to describe more than a single type of MOS device. However, all MOSFETs behave similarly to the enhancement-mode

[1]Continuing increase of V_{DS} gives rise to a breakdown effect, wherein the current again sharply increases with V_{DS}. Operation of the device in breakdown is undesirable and avoided in practice. Breakdown is not considered in these models or this report.

transistor presented here, from a modeling point of view.

2.3.2. Output Characteristics A first-order analytical MOS transistor model that follows from a simple charge-control analysis is presented in this subsection. The model is used to illustrate several points later in this report. An n-channel enhancement-mode device is assumed.

When all terminals are connected to ground no conducting channel is present. If the gate voltage is increased positively from zero with respect to the source, negative charge is attracted to the surface of the silicon at the $Si-SiO_2$ interface. Inversion occurs, and a conducting channel is formed, when V_{GS} reaches the threshold voltage V_T. Above threshold, the number of additional electrons attracted to the surface is proportional to the gate-source voltage V_{GS}. The added electrons are readily supplied by the nearby n-type source and drain regions.

When the substrate terminal is negatively biased with respect to the source, i.e., $V_{BS} < 0$, a larger V_{GS} value is required to reach threshold than for $V_{BS} = 0$. This phenomenon is called the *body effect*. A first-order analysis, such as that in [8] or [9], shows that

$$V_T = V_{TO} + \gamma\left[\sqrt{2|\varphi_F| - V_{BS}} - \sqrt{2|\varphi_F|}\right]. \tag{2.7}$$

The parameter V_{TO} in Eq.(2.7) is the threshold voltage for $V_{BS} = 0$, φ_F is the equilibrium potential at the Si surface, and γ depends on material parameters and the gate oxide thickness.

If V_{GS} is above threshold, and if a small positive voltage is applied between the drain and source (V_{DS}), an electron current flows from drain to source due to drift. The drain-to-source current I_{DS} (or more simply, the drain current) is related to the channel charge and the transit time τ_c along the channel. As long as a continuous channel exists from the source to the

drain, the current is given by [10]

$$I_{DS} = \frac{\mu_n C_{ox} W}{L}\left[V_{GS} - V_T - \frac{V_{DS}}{2}\right]V_{DS} \quad . \tag{2.8}$$

Equation (2.8) predicts the drain current for the so-called *linear* (also known as triode or resistance) region. The linear region corresponds to $V_{GS} > V_T$ and $V_{DS} < V_{GS} - V_T$ for this model. If the drain-to-source voltage is increased above $V_{GS} - V_T$, *pinch-off* occurs and the MOSFET enters the so-called *saturation* region.

The onset of saturation occurs when V_{DS} becomes so large that the channel no longer extends from the source to the drain, but instead stops at the point L' in Figure 2.3. In saturation, the voltage difference between the gate and the channel is less than V_T in the region near the drain p-n junction. The channel is thus depleted of carriers there and the electrical channel length is reduced to L'. When the electrons reach L', they are quickly swept across the depleted area by the large electric field between L' and L, and I_{DS} is relatively insensitive to further increases in V_{DS}.

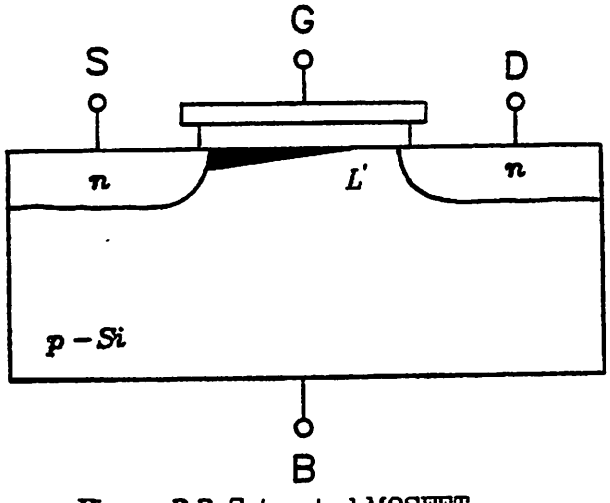

Figure 2.3 Saturated MOSFET.

For this simple model, the current in saturation is found by replacing V_{DS} by $V_{GS} - V_T$ in Eq.(2.8) yielding

$$I_{DS} = \frac{\mu_n C_{ox} W}{2L}\left[V_{GS} - V_T\right]^2 . \tag{2.9}$$

Equation (2.9) is valid for $V_{GS} > V_T$ and $V_{DS} \geq V_{GS} - V_T$. Most MOS transistors have output characteristics that vary slightly with V_{DS} in saturation, as in Figure 2.4. A representation of this nonzero saturation-region output conductance can be included by changing Eq.(2.9) to [10]

$$I_{DS} = \frac{\mu_n C_{ox} W}{2L}\left[V_{GS} - V_T\right]^2(1 + \lambda V_{DS}). \tag{2.10}$$

To maintain current and partial derivative continuity, Eq.(2.8) must have this term added also and thus becomes

$$I_{DS} = \frac{\mu_n C_{ox} W}{L}\left[V_{GS} - V_T - \frac{V_{DS}}{2}\right]V_{DS}(1 + \lambda V_{DS}). \tag{2.11}$$

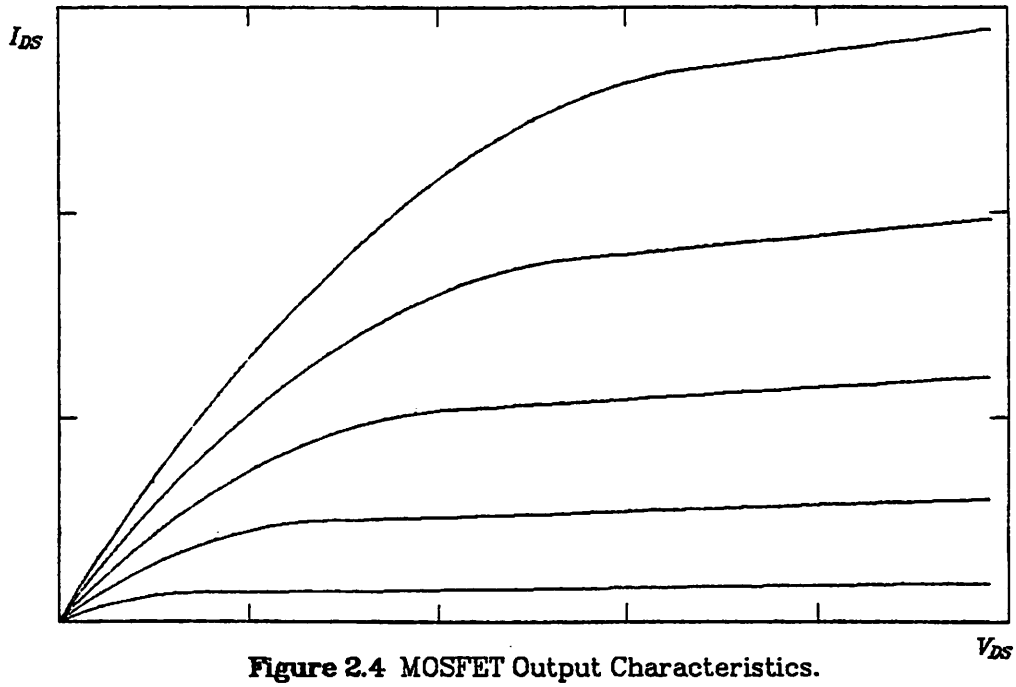

Figure 2.4 MOSFET Output Characteristics.

An examination of this simple model illuminates some important aspects of MOSFET behavior. These are:

A. The output characteristics vary strongly with V_{DS} in the linear region, but weakly in the saturation region.

B. The body bias V_{BS} appears only in the threshold voltage equation.

C. Given the current at the point of saturation, little additional information is needed to model the output characteristics in the saturation region.

Items A through C will prove to be useful later in the development of the empirical models.

2.4. MOSFET REPRESENTATION FOR SIMULATION

The NR method is used as a *rule* in this section to derive the MOSFET dc companion model. The simple analytical model from Section 2.3 is utilized as an example.

The equation to be linearized has the form

$$I_{DS} = f(V_{DS}, V_{GS}, V_{BS}) \qquad (2.12)$$
$$= f(\bar{V}).$$

Equation (2.2) through the linear term produces

$$f(\bar{V}) = f(\bar{V}_0) + (\bar{V} - \bar{V}_0) f'(\bar{V}_0),$$

and since $\bar{V}_0$ is the value of $\bar{V}$ at the j^{th} iterate,

$$f(\bar{V}_{j+1}) = f(\bar{V}_j) + f'(\bar{V}_j)\bar{V}_{j+1} - f'(\bar{V}_j)\bar{V}_j. \qquad (2.13)$$

Equations (2.12) and (2.13) are combined and rearranged to give

$$I_{DS_{j+1}} = \left[\frac{\partial f}{\partial V_{DS}} \quad \frac{\partial f}{\partial V_{GS}} \quad \frac{\partial f}{\partial V_{BS}}\right]_j \begin{bmatrix} V_{DS} \\ V_{GS} \\ V_{BS} \end{bmatrix}_{j+1} +$$

$$\left\{ f(V_{DS}, V_{GS}, V_{BS})_j - \left[\frac{\partial f}{\partial V_{DS}} \quad \frac{\partial f}{\partial V_{GS}} \quad \frac{\partial f}{\partial V_{BS}}\right]_j \begin{bmatrix} V_{DS} \\ V_{GS} \\ V_{BS} \end{bmatrix}_j \right\}. \quad (2.14)$$

The term of Eq.(2.14) in braces is composed entirely of quantities from the j^{th} iteration. It has the dimensions of current, and is denoted I_{EQ_j}. Equation (2.14) can thus be rewritten as

$$I_{DS_{j+1}} = I_{EQ_j} + \left[\frac{\partial f}{\partial V_{DS}}\right]_j V_{DS_{j+1}} + \left[\frac{\partial f}{\partial V_{GS}}\right]_j V_{GS_{j+1}} + \left[\frac{\partial f}{\partial V_{BS}}\right]_j V_{BS_{j+1}} . \quad (2.15)$$

Equation (2.15) defines the MOSFET companion model, which is pictured in Figure 2.5. The partial derivative terms have the dimensions of conduc-

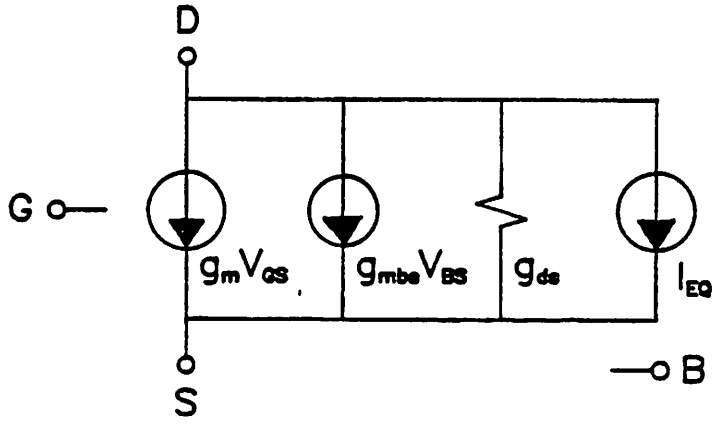

Figure 2.5 MOSFET Companion Model.

tance, and are commonly symbolized and named as presented in Table 2.1.

Term	Symbol	Name
$\frac{\partial f}{\partial V_{DS}} = \frac{\partial I_{DS}}{\partial V_{DS}}$	g_{ds}	*output conductance*
$\frac{\partial f}{\partial V_{GS}} = \frac{\partial I_{DS}}{\partial V_{GS}}$	g_m	*transconductance*
$\frac{\partial f}{\partial V_{BS}} = \frac{\partial I_{DS}}{\partial V_{BS}}$	g_{mbs}	*backgate conductance*

Table 2.1 Partial Derivatives.

The various conductances are determined from the drain current equation. As an example, suppose that Eq.(2.9) applies. In that case, g_m is found to be

$$g_m = \frac{\partial I_{DS}}{\partial V_{GS}} = \frac{\partial}{\partial V_{GS}}\left\{\frac{\mu_n C_{ox} W}{2L}\left[V_{GS} - V_T\right]^2\right\}$$

$$= \frac{\mu_n C_{ox} W}{L}\left[V_{GS} - V_T\right].$$

The other two conductances are found similarly.

In the following chapter, empirical and analytical methods of determining the element values for the companion model are described.

CHAPTER 3

MOSFET MODELING METHODS

3.1. INTRODUCTION

In Chapter 2, some of the general techniques used in modeling nonlinear devices for circuit simulation are described. It is shown that the problem amounts to providing a proper incremental circuit model for the device, which corresponds to a linearization of its $I-V$ characteristics about a given operating point. Any MOSFET model should return element values efficiently and accurately for the incremental circuit model. The element values returned by the model must satisfy the convergence constraints imposed by the Newton-Raphson algorithm.

Various analytical and semi-empirical models offer increasing accuracy at the expense of decreasing computational efficiency. *Empirical* models for circuit simulators exhibit somewhat different compromises when higher accuracy is desired. The amount of memory needed for storing data, the complexity of the interpolation procedures, and the accuracy of the model can all be exchanged with one another as desired.

In this chapter, several analytical MOSFET models are briefly described. The difficulties with analytical modeling are presented, followed by the means by which a good empirical model could solve most of the problems of analytical models. Some early work in the area of empirical MOSFET modeling is outlined, as well as some of the recent research in the field. The chapter concludes with a section on the basics of the approach to empirical modeling taken in this project.

3.2. ANALYTICAL MODELS

This section on analytical models is included as background for a later comparison with empirical models.

The MOSFET models built into the circuit simulation program SPICE2 [11] illustrate the level of complexity accurate models possess. Two models in SPICE2 are appropriate for modern small-geometry MOS transistors, the analytical LEVEL-2 model and the semi-empirical LEVEL-3 model. Comprehensive explanations of these models appear in [1] and [12].

The LEVEL-2 and LEVEL-3 models account for various nonideal effects, such as scattering-limited carrier drift velocity, field-dependent mobility, etc. On the order of 10 parameter values must be supplied to specify the dc behavior of the models [12]. As noted in [1], a means of generating model parameter values is as critical as the fundamental accuracy of a model. Unfortunately, the parameter values for good-quality models such as these are difficult to determine, because relatively few values are forced to contain a large amount of compressed information. Often, a procedure for calculating the parameter values does not exist, and the task must be performed manually using initial guesswork and trial and error.

The completeness of analytical and semi-empirical models like LEVEL-2 and LEVEL-3 leads to the conclusion that the time required to evaluate the model equations is relatively large.[1] As stated in Chapter 2, model evaluation can dominate the simulation time for some circuits. Compared to the SPICE2 LEVEL-1 model, which is basically the first-order model described in Chapter 2, the LEVEL-2 model is about 16 times less efficient [13] to evaluate. The LEVEL-3 model is up to 40% more efficient than LEVEL-2 [1], but is still a

[1]The term analytical includes semi-empirical for the remainder of this chapter.

factor of nearly 10 less efficient than LEVEL-1.

As a particular process or technology evolves, the corresponding analytical model must usually be changed. For example, suppose a channel implant is added to an existing process. To model the modified process, terms typically must be added to the model equations, or perhaps the model must be completely reformulated.

The inherent time lag between a technological innovation and its understanding on the device physics level can causes two problems. First, the necessarily outdated analytical model may not be capable of fitting the new device. Second, the process and/or technology-dependent model parameters must be used to curve-fit the new device, since the physics of the new device are not fully incorporated in the model. The model parameters then lose some of the physical significance they originally had, leading to possible confusion and error on the part of the model user.

Analytical models are typically scaled with channel length and channel width. The range over which a given model can be scaled is usually limited, though. Several models are often needed, each valid for a part of the complete range of sizes.

3.3. EMPIRICAL MODELS AND THEIR ADVANTAGES

Some of the general characteristics of empirical device models are presented in this section. A description of the advantages empirical models have over analytical models is included also.

Empirical models do not contain any process or technology-dependent parameters. The absence of parameters dependent on the process or the technology eliminates the (difficult) step of determining their values.

The data set (tables of $I-V$ data or coefficients of fitting functions) of an empirical model can be made very general and dependent only on the *most basic* aspects of the device's behavior. If this aim is achieved, the empirical model will not require revision for changes in the process, geometry scaling, etc. Rather, all that is necessary to account for an additional process step or other change is a regeneration of the table entries or fitting-function coefficients. Maintaining generality in this fashion places a lower limit on the size of an empirical model data set.

The data set required for an empirical model cannot be prohibitively large. An excessively large data set may preclude the use of many different empirical models in a single circuit simulation, due to computer memory limitations. It is desirable to store a distinct model for each device type and each channel length.[2] If only a small number of models can be accommodated, then they must be scaled by channel length as well as width. The scaling of MOSFETs by channel length is often highly nonlinear and should be avoided.

The generation of the data set for an empirical model is performed much more easily than parameter value determination for an analytical model.[3] Empirical models have this advantage because the data set for an empirical model is significantly larger than the parameter set for an analytical model. Less information compression is then required to characterize the empirical model. A secondary benefit to a large data set compared to a small parameter set lies in the area of sensitivity. A small error in a critical parameter value for an analytical model results in a much greater

[2] In the design of many MOS ICs, channel length is restricted to a few discreet sizes, but channel width is allowed to vary continuously.

[3] The appendix of this report contains a brief description of the automatic characterization procedure which has been implemented for the models of Chapter 4 and Chapter 5.

perturbation in the model's accuracy, compared to the same magnitude of error in a value in an empirical model's data set.

Since they are simple to characterize, empirical models have the benefit of fitting flexibility. Figure 3.1 shows this pictorially. Device simulation can provide raw data for the empirical models. Direct measurement of devices, perhaps using automated test equipment, is another possibility. A third is the use of an existing analytical model and its parameter values. This alternative is useful when the empirical model possesses a speed advantage over

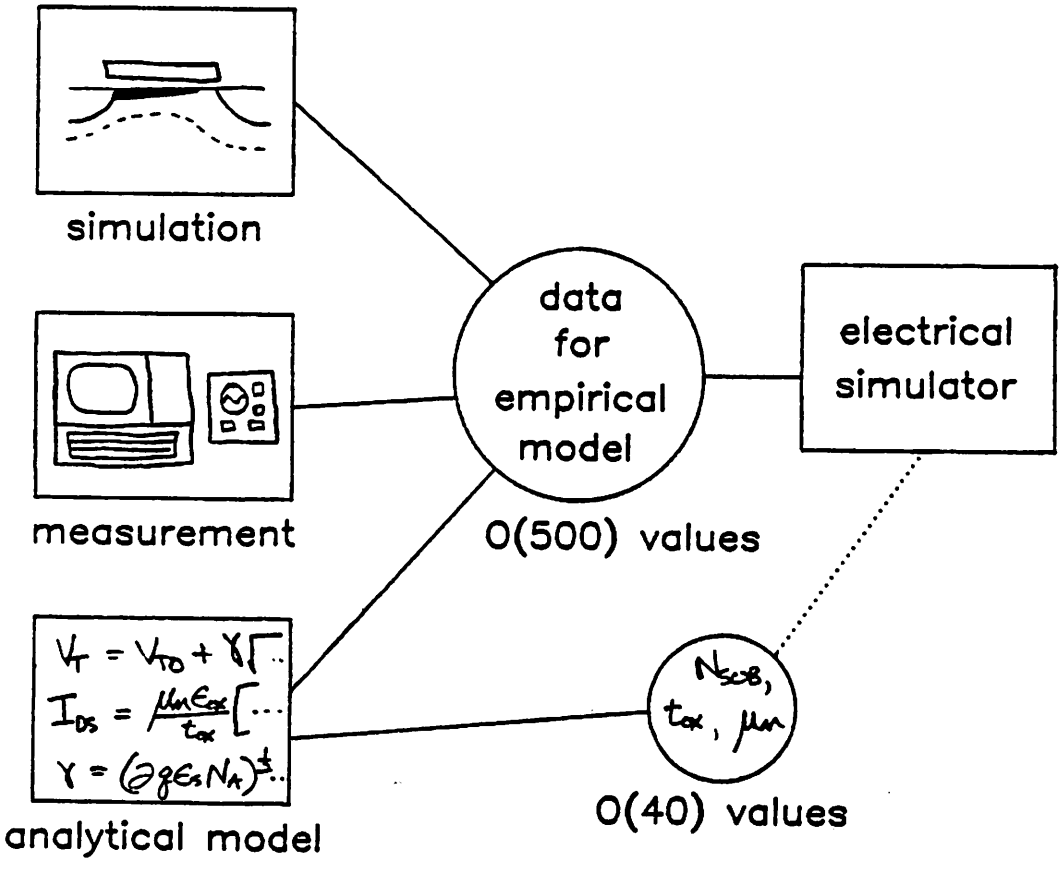

Figure 3.1 Sources of Empirical Model Data

the analytical model.

Empirical models are capable of producing an arbitrarily accurate reproduction of the $I-V$ characteristics which were used to generate the model data set, by increasing the size of the data set. Analytical models do not have this property.

Finally, an empirical model can be computationally efficient. Practical empirical models, such as the two described in this report, are significantly faster to evaluate than comparably accurate analytical models, because only simple arithmetic operations and memory references are used to generate the drain current and conductance values for the companion model.

3.4. EMPIRICAL MODEL BASICS

In this section, some general observations regarding empirical MOSFET models are outlined. Then, two basic types of empirical models are described, table look-up models and function-fit models. The empirical models developed in this research employ both look-up tables and function fits. The principles underlying their development are presented in the following section.

The structure of an empirical model involves consideration of three competing goals. These are the desire for high accuracy in reproducing the $I-V$ behavior, the desire for low data storage requirements, and the desire for computational efficiency. In this work, high accuracy has been a primary concern. The first problem to be faced, then, is to obtain the accuracy of a good-quality analytical model within a reasonable storage allocation for the model data set. The second is to achieve an advantage in evaluation speed over a good-quality analytical model.

The MOSFET companion model, as noted in Chapter 2, comprises four elements. The MOSFET itself has four terminals; hence three independent node-pair voltages determine the drain current and the partial derivatives. A consequence of these two facts is that knowledge of four three-dimensional quantities is required, namely

$$I_{DS} = I_{DS}(V_{DS}, V_{GS}, V_{BS}), \tag{3.1a}$$

$$g_{ds} = g_{ds}(V_{DS}, V_{GS}, V_{BS}), \tag{3.1b}$$

$$g_m = g_m(V_{DS}, V_{GS}, V_{BS}), \tag{3.1c}$$

$$g_{mbs} = g_{mbs}(V_{DS}, V_{GS}, V_{BS}). \tag{3.1d}$$

The problem to be solved is that of storing information from which the I_{DS}, g_{ds}, etc. values can be reproduced accurately and efficiently at a specific operating point. A general comment can be made regarding Eqs.(3.1); because of their three-dimensional nature, either function-fit or table look-up models need large amounts of storage if the physics of the device are not accounted for in the model structure.

3.4.1. Table Look-Up Models A straightforward approach to storing Eqs.(3.1) is to place each one in a table. Each table entry, e.g., each I_{DS} value, is indexed by a (V_{DS}, V_{GS}, V_{BS}) triplet, so the tables are *three-dimensional.* At a particular operating point, the values for I_{DS} and the conductances are "looked-up" in the tables.

A major point in favor of table look-up models is their conceptual simplicity. However, close examination shows that several difficult problems exist for table look-up models. Table look-up methods typically require a minimum of 50-100 points per dimension to insure an adequate fit [14]. If 50 points per dimension is assumed sufficient, Eqs.(3.1) require storage for $4(50)^3 = 500{,}000$ values. This number of required memory locations is

impractically large.

Another problem exists for this method. The drain current and the partial derivatives are explicitly stored for discreet operating points only. During a circuit simulation, operating points arise that do not coincide with the explicitly-stored data. The element values for the companion model must then be computed from the stored data via an interpolation method. The interpolations are in general three-dimensional and nonlinear, and hence are computationally cumbersome. If instead a sequence of single-dimensional interpolations is used, and/or if the interpolations are done linearly, accuracy suffers and continuity cannot be guaranteed.

Recent research by Shima et al. [15] has resulted in a three-dimensional table look-up model which has more favorable storage requirements than the three-dimensional method outlined here. The authors were able to reduce the storage to 2000-3000 points per model, primarily by computing the conductances from the I_{DS} table and eliminating the g_{ds}, g_m, and g_{mbs} tables, and partially by storing less data in the V_{BS} dimension than in the V_{DS} or V_{GS} dimensions in the I_{DS} table. However, the interpolation routines do not guarantee current continuity, and the partial derivatives which result are also discontinuous functions. It is therefore impossible to insure convergence using this model.

Other table look-up models have been used in the past, but typically not for circuit simulation. The models [16], [17] have been used in a type of timing simulator where model requirements are much less stringent. These models do not generate g_m and g_{mbs}. Also, the I_{DS} and g_{ds} functions produced are discontinuous, because interpolation is not used. Models such as these are not applicable to circuit simulation.

3.4.2. Function-Fit Models A second method exists for storing Eqs.(3.1). Numerical functions with no specific relation to the device electronics can be fit to the $I-V$ characteristics. At an operating point, the functions are evaluated and return I_{DS}, g_{ds}, etc.

Certain fitting functions have convenient properties from the standpoint of the NR method. For example, a quadratic or cubic spline fit of Eq.(3.1a) works well, because then the remaining parts of Eqs.(3.1) can be generated by explicit partial differentiation of Eq.(3.1a). Continuous first partial derivatives for quadratic or higher-order splines are guaranteed by the definition of the spline function [18].

A negative aspect of spline models is that the requisite storage can be very high. Approximately four times the number of points to be fit is required for storing the coefficients for a one-dimensional cubic spline, and nearly three times the number of points to be fit is required for storing the coefficients for a one-dimensional quadratic spline. The number of coefficients required per point fit is larger still for splines of more than one dimension.

Recent work in spline models for MOSFETs [19] has produced encouraging results in model evaluation speed. However, the storage versus accuracy issue could be a problem; the presentation in [19] does not explore this question.

3.5. BASICS OF THIS APPROACH TO EMPIRICAL MODELING

The two empirical models of this project have been developed using several ideas in common. These concepts illustrate why the models evolved to their present states. Each of the two models is described in detail in the next two chapters of this report.

The general, the global behavior of MOSFETs is well known, although the local behavior can be quite variable. The MOSFET behavior is exploited with the aim of developing more optimal models. The first concept is thus to *use the known physical behavior to reduce the storage requirement.* Physical knowledge is used to decouple the independent variables so that the data set can be stored in two-dimensional and one-dimensional functions instead of three-dimensional functions. *Which* information is stored is also chosen after consideration of the device physics. Specifically,

A. The V_{BS} dependence is assumed to affect only the effective gate-source voltage.

B. The output characteristic data is explicitly stored for the linear region only.

C. The output characteristic data is implicitly stored for the saturation region, using single-dimensional functions.

D. As in [15] and [19], the conductance terms are computed from the $I-V$ information and not explicitly stored.

The second idea followed in developing the models is to *use a combination of fitting functions and tables* for the data set storage. Tables are used to store the $I-V$ data where the curvature of the characteristics is high and much information is present. Splines would be too storage-intensive in such areas, as noted in the preceding section. A problem resulting from tabulating the data is maintaining continuity of the partial derivatives. But, the amount of data stored can usually be increased to the point where the discontinuity of the partial derivatives does not affect the convergence of the simulation program without reaching the amount of storage required for spline coefficients. Spline functions are used in regions where the data is known to

be smooth and slowly-varying. In these regions of the characteristics, there is often less storage required for the spline coefficients than for tabulation of the same data under the condition that the partials be well-behaved. In particular, single-dimensional functions are used at the linear-to-saturation transition to insure current continuity and well-defined partial derivatives there.

The third principle followed in this work is to *use knowledge of the device physics to simplify the look-up table interpolations*. Once the independent variables are decoupled, the dimensionality of the interpolations is correspondingly reduced. The functions used for the interpolations are chosed after consideration of the device physics, *in that region of operation*. If this were not done, more complex interpolation functions would be required and/or more data would have to be stored.

Detailed descriptions of the two-dimensional and one-dimensional models are presented in the following two chapters.

CHAPTER 4

TWO-DIMENSIONAL EMPIRICAL MODEL

4.1. INTRODUCTION

The previous chapter describes the major problem inherent in a three-dimensional empirical model, namely that the 3-d approach is quite memory-intensive. The one-dimensional model of Chapter 5 requires only a small storage allocation. However, too much information *may* be lost in the reduction of the $I-V$ characteristics to strictly 1-d functions. The two-dimensional empirical model described in this chapter falls between the 1-d and 3-d methods in terms of potential accuracy and required memory.

The reduction of the data set from dimensionality of three to dimensionality of two is achieved by decoupling the V_{BS} dependence from the $I-V$ characteristics. A further memory reduction follows from storing data as a function of two independent variables for the linear region only. The remainder of the $I-V$ information is stored in one-dimensional functions.

The following sections describe the two-dimensional model in detail.

4.2. MODEL DESCRIPTION

The bulk-source potential of a MOSFET has a large influence on the $I-V$ characteristics through the threshold voltage V_T, but only a small influence otherwise [15]. Furthermore, V_{GS} never appears alone, but always has V_T associated with it as an offset term. These properties lead one to store the characteristics using $V_{GSE} = V_{GS} - V_{TO}$ as the independent variable in place of

V_{GS}. Then, the effect of V_{BS} on V_T can be stored in a one-dimensional function defined as

$$V_T = S_1(V_{BS}).^1$$

The drain current is thereby reduced to a two-dimensional quantity plus a one-dimensional function:

$$I_{DS}(V_{DS}, V_{GS}, V_{BS}) \rightarrow I_{DS}(V_{DS}, V_{GSE}) \text{ with } V_T = S_1(V_{BS}).$$

The main feature of the 2-d model is the drain current table. A two-dimensional table is used, with axes corresponding to V_{DS} and V_{GSE}. As noted in Chapter 2, MOSFET output characteristics vary strongly with V_{DS} in the linear region but weakly in saturation. The data is therefore stored for the linear region only. The output characteristics are stored as discreet points as shown in Figure 4.1.

The linear region is demarcated by the V_{DS} values (for each V_{GSE}) where the MOSFET enters the saturation region. The drain current and drain-

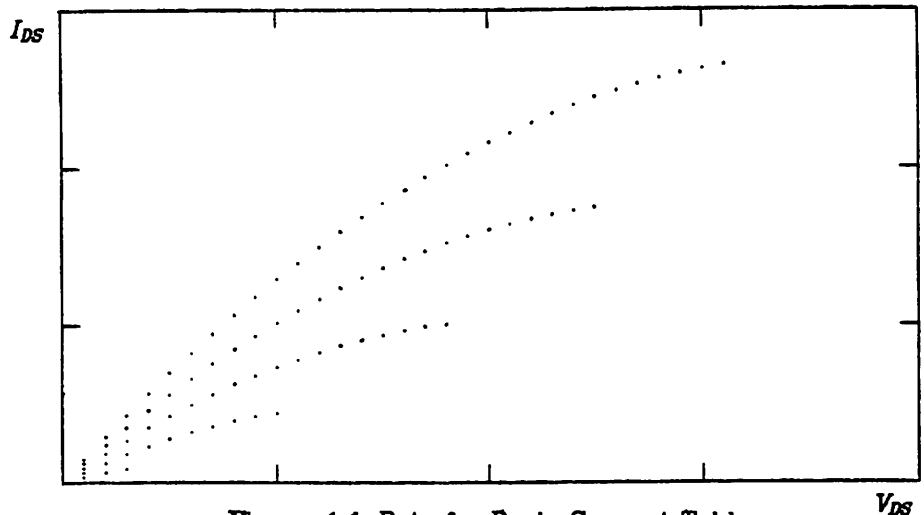

Figure 4.1 Data for Drain Current Table.

[1] The S_i notation means the function is a cubic spline.

source voltage where this occurs are denoted I_{DSAT} and V_{DSAT}, respectively. For the 2-d model, the saturation point is defined as that point where the output conductance can be approximated as a constant.[2] The $I_{DSAT}-V_{DSAT}$ curve is stored as a cubic spline function.

A cubic spline relating V_{DSAT} to V_{GSE} is used to complete the description of the characteristics at the linear-to-saturation transition. Splines are used for the $I_{DSAT}-V_{DSAT}$ and $V_{DSAT}-V_{GSE}$ variations because these variations are smooth but variable. For example, a long-channel transistor has $V_{DSAT} = V_{GSE}$, and $I_{DSAT} \alpha V_{DSAT}^2 (= V_{GSE}^2)$ [10]. A short-channel device, however, typically has an $I_{DSAT}-V_{GSE}$ relation at the saturation point which is close to linear [21]. Complicating the issue is the fact that some MOSFETs show long-channel behavior (in this respect) for small V_{GSE} values but then show short-channel behavior for larger V_{GSE} values. The spline functions give smooth characteristics and smooth partial derivatives, while allowing the behavior to be general.

To complete the model the output characteristics in saturation must be stored. The output conductance of a MOSFET is well-approximated by linear equations in V_{DS} with slopes which increase as V_{GSE} increases (Figure 4.2). The slope of the $I_{DS}-V_{DS}$ characteristics, i.e., g_{ds}, is stored indirectly through the incorporation of a fourth cubic spline, $I_{DMAX} = S_4(V_{GSE})$. The S_4 spline is fit at a large V_{DS} value called V_{DMAX}. Along with the $I_{DSAT}-V_{DSAT}$ and $V_{DSAT}-V_{GSE}$ curves, S_4 provides a means for generating I_{DS}, g_{ds}, g_m, and g_{mbs} for operating points in the saturation region.

[2] The procedure used for determining the saturation point is outlined in the appendix.

In summary, the 2-d model's data set is:

$$I_{DS_{lin}} = T_{DS}(V_{DS}, V_{GSE})$$

$$V_T = S_1(V_{BS})$$

$$V_{DSAT} = S_2(V_{GSE})$$

$$I_{DSAT} = S_3(V_{DSAT})$$

$$I_{DMAX} = S_4(V_{GSE}),$$

where

$$V_{GSE} = V_{GS} - V_T$$

and T_{DS} denotes the two-dimensional table for the linear-region characteristics.

The 2-d model has dramatically reduced requisite storage over the simple 3-d model. Each cubic spline requires $4(k-1)$ coefficients if k data points are fit [7]. The T_{DS} table contains less than n^2 values if n $V_{GSE} = const.$ curves are stored with n points in V_{DS} on the curve where V_{GSE}

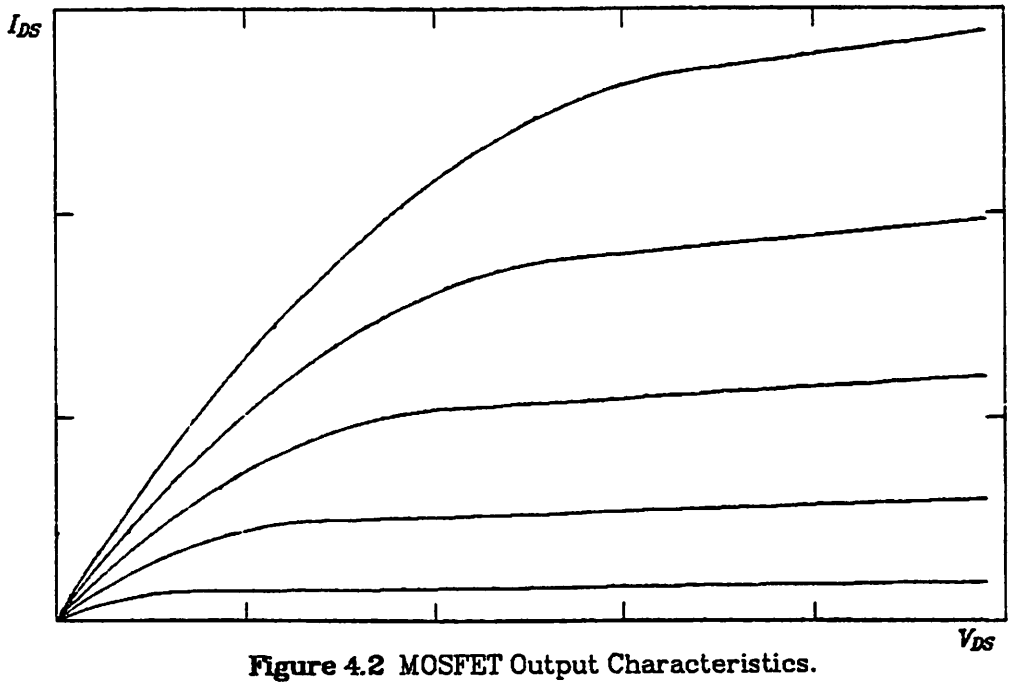

Figure 4.2 MOSFET Output Characteristics.

is largest. Figure 4.1 shows this fact pictorially. The total memory requirement for the 2-d model is thus less than $n^2 + 16(k-1)$ locations.

The threshold voltage is assumed to vary only with V_{BS} in the above. Many short-channel MOSFETs exhibit a V_T variation with V_{DS} also. The variation of V_T with V_{DS} can be modeled as an additive effect [22], and if needed a fifth spline can be incorporated for the V_{DS} dependence. The threshold voltage equation would then read

$$V_T = S_1(V_{BS}) + S_5(V_{DS}).$$

4.3. CURRENT CALCULATION

Due to the manner in which the data set is stored, the drain current and partial derivative calculations are *operating-region dependent* for the 2-d empirical model. Three different cases exist; each is addressed below. The precursory step of determining the region in which the operating point lies is performed using the S_1 and S_2 splines:

$V_{GSE} = V_{GS} - S_1(V_{BS})$;

$V_{DSAT} = S_2(V_{GSE})$;

if ($V_{DS} \geq V_{DSAT}$) then

transistor is in saturation;

else

transistor is in linear.

4.3.1. Case 1: Simple Linear This case occurs for operating points that lie in the linear region where the two V_{GSE} = *const.* characteristics bounding the operating-point V_{GSE} extend *beyond* the operating-point V_{DS}. The situation is depicted in Figure 4.3.

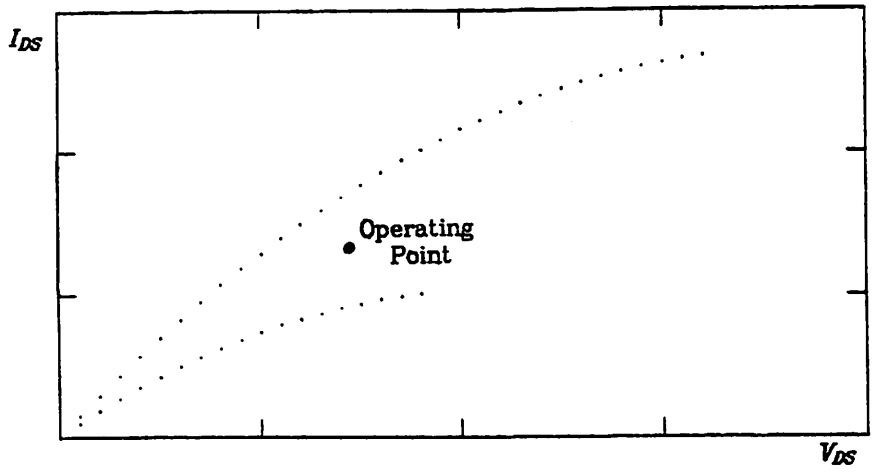

Figure 4.3 Simple Linear Operating Point.

Two steps are required to find I_{DS}. Figure 4.4 displays the I_{DS} calculation. Initially, quadratic interpolation is used along the V_{GSE_1} and V_{GSE_2} curves in the V_{DS} dimension to calculate I_{DS_1} and I_{DS_2}. Then, these intermediate drain current values are interpolated linearly in V_{GSE} to determine I_{DS}. Thus, the drain current calculation for this case uses only the data in

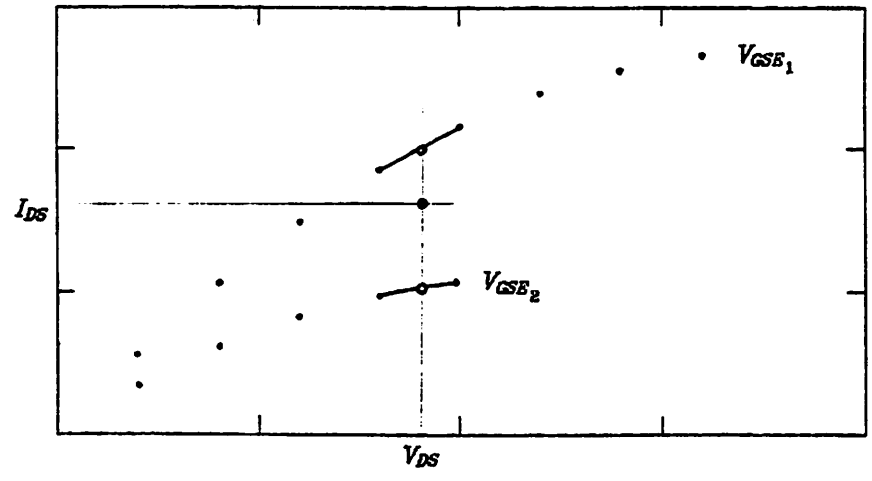

Figure 4.4 Simple Linear Current Calculation.

the T_{DS} table, once V_{GSE} is known.

The output conductance, g_{ds}, is determined concurrently with I_{DS} (Figure 4.5). The derivative of the quadratic interpolation function used along the V_{GSE_1} and V_{GSE_2} curves provides intermediate conductances g_{ds_1} and g_{ds_2}, respectively. The final value for g_{ds} follows from a linear interpolation of g_{ds_1} and g_{ds_2} in the V_{GSE} dimension.

The transconductance, g_m, is simply the derivative of the linear interpolation function used to find I_{DS} from I_{DS_1} and I_{DS_2}. The backgate conductance, g_{mbs}, follows from the derivative of the S_1 spline, g_m, and the chain rule. Symbolically,

$$g_{mbs} = \frac{\partial I_{DS}}{\partial V_{BS}} = \frac{\partial I_{DS}}{\partial V_{GSE}} \frac{\partial V_{GSE}}{\partial V_{BS}} = g_m \left[\frac{-dS_1}{dV_{BS}} \right].$$

4.3.2. Case 2: Out-of-Bounds Linear In this case the operating point lies in the linear region. However, data is not available on the V_{GSE_1} contour at the operating-point V_{DS}. Figure 4.6 shows the problem. This case is the most

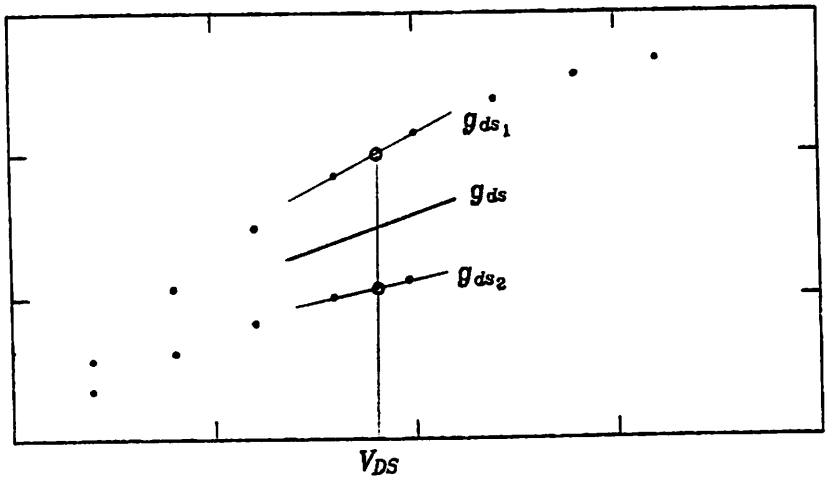

Figure 4.5 Simple Linear Output Conductance Calculation.

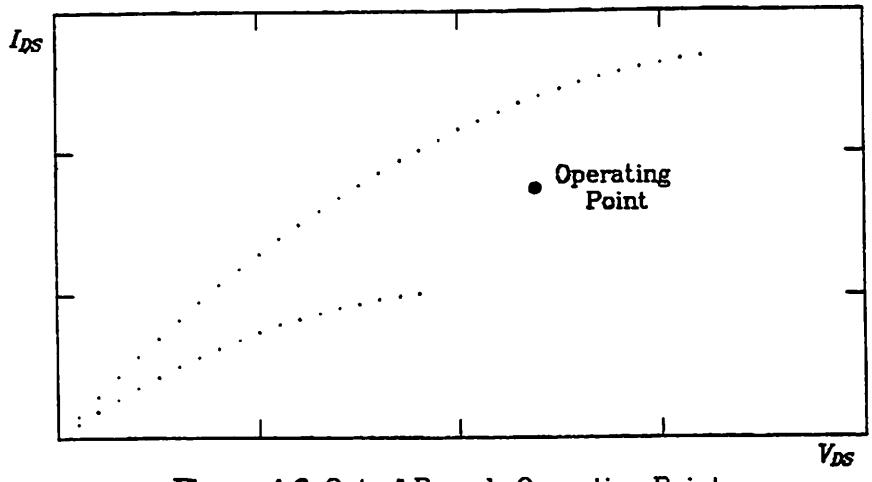

Figure 4.6 Out-of-Bounds Operating Point.

difficult of the three, from the standpoint of maintaining accuracy and partial derivative continuity.

The first step in calculating I_{DS} is to find I_{DSAT} at V_{GSE}. This is done via the spline function which relates I_{DSAT} and V_{DSAT}, i.e.,

$$I_{DSAT} = S_3(V_{DSAT}) = S_3(S_2(V_{GSE})) .$$

Second, the current at V_{GSE} and $V_{DS} = V_{DSAT_1}$ is found by a linear interpolation in V_{GSE} using two values from T_{DS}, as depicted in Figure 4.7a. The final I_{DS} value is found by a quadratic interpolation in V_{DS} using I_{DSAT} and $I_{DS}(V_{DSAT_1}, V_{GSE})$, Figure 4.7b.

The value for g_{ds} is the derivative of the quadratic interpolation function used between I_{DSAT} and $I_{DS}(V_{DSAT_1}, V_{GSE})$, evaluated at the operating-point V_{DS}. Figure 4.7b also shows g_{ds}.

The transconductance is more difficult to calculate than the output conductance. At V_{DSAT},

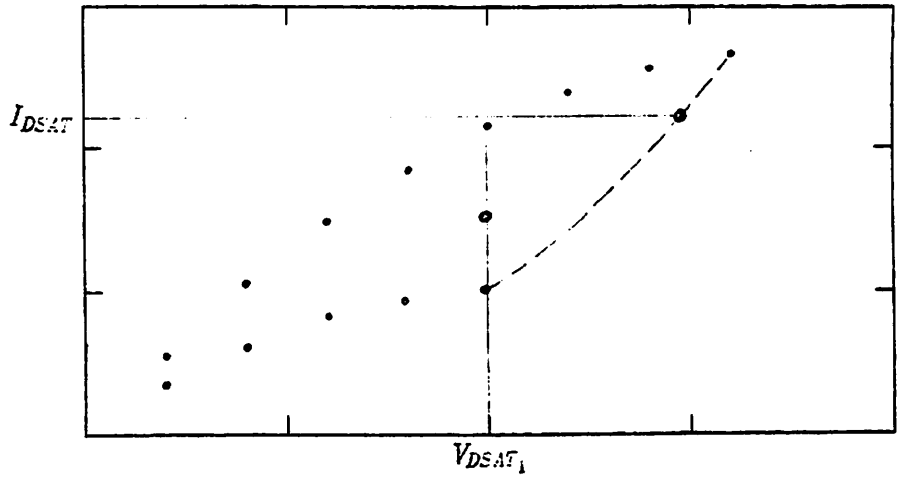

Figure 4.7a Out-of-Bounds Calculation.

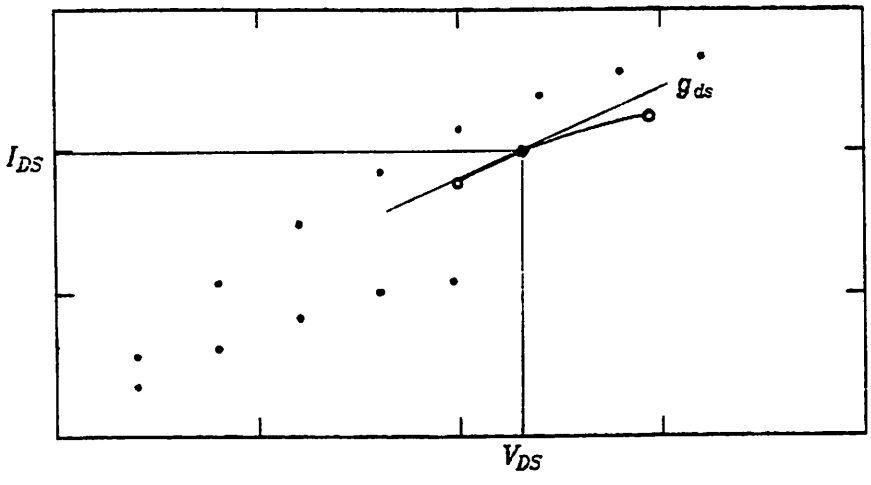

Figure 4.7b Out-of-Bounds Calculation.

$$g_m = g_{m_{sat}} = \frac{dI_{DSAT}}{dV_{DSAT}} \frac{dV_{DSAT}}{dV_{GSE}}$$

$$= \frac{dS_3}{dV_{DSAT}} \frac{dS_2}{dV_{GSE}}$$

At $I_{DS}(V_{DSAT_1}, V_{GSE})$, the derivative of the linear interpolation function in the V_{GSE} direction gives $g_m = g_{m_{lin}}$. The transconductance at the operating point

is calculated via linear interpolation of $g_{m_{lin}}$ and $g_{m_{sat}}$ in V_{DS}.[3] The backgate conductance g_{mbs} is given by

$$g_{mbs} = g_m \frac{-dS_1}{dV_{BS}} .$$

4.3.3. Case 3: Saturation The drain current and the partial derivatives are easily found for operating points in the saturation region. Since V_{DSAT} is known, I_{DSAT} follows from $I_{DSAT} = S_3(V_{DSAT})$. At $V_{DS} = V_{DMAX}$, $I_{DMAX} = S_4(V_{GSE})$ is calculated. The drain current results from

$$I_{DS} = I_{DSAT} + \left(\frac{I_{DMAX} - I_{DSAT}}{V_{DMAX} - V_{DSAT}} \right) V_{DS} .$$

The term in parenthesis in the above equation is the output conductance, i.e.,

$$g_{ds} = \left(\frac{I_{DMAX} - I_{DSAT}}{V_{DMAX} - V_{DSAT}} \right) .$$

Figure 4.8 contains a pictorial representation of the I_{DS} and g_{ds} determinations. The transconductance is found as follows from splines S_2, S_3, and S_4. At V_{DMAX},

$$g_{m_{max}} = \frac{dI_{DMAX}}{dV_{GSE}}$$

$$= \frac{dS_4}{dV_{GSE}}$$

At the saturation point, as in Case 2 above,

$$g_{m_{sat}} = \frac{dI_{DSAT}}{dV_{DSAT}} \frac{dV_{DSAT}}{dV_{GSE}} .$$

[3] Actually, g_m at the saturation point is slightly in error because it is taken as the partial derivative of I_{DSAT} rather than I_{DS}.

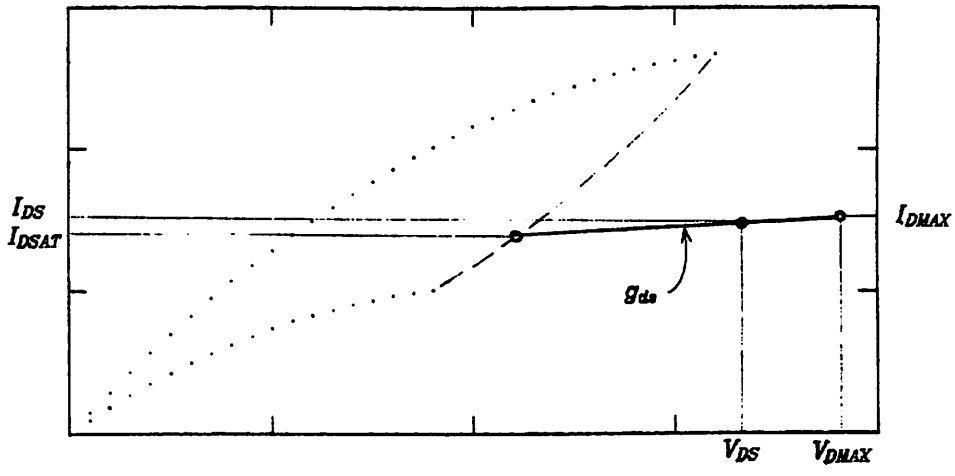

Figure 4.8 Saturation Current and Output Conductance.

A linear interpolation in V_{DS} generates the g_m value at the operating point, i.e.,

$$g_m = g_{m_{sat}} + \left\{ \frac{g_{m_{max}} - g_{m_{sat}}}{V_{DMAX} - V_{DSAT}} \right\} V_{DS}.$$

The backgate conductance is calculated from g_m as in Case 1 and Case 2.

4.4. CONTINUITY CONSIDERATIONS

The structure of the two-dimensional model guarantees drain current continuity. However, the same statement cannot be made for the partial derivatives in the linear region. These points are covered in this section, as are means for reducing the risk of nonconvergence of the NR method.

4.4.1. Linear Region In the V_{DS} dimension, all the interpolations are along $V_{GSE} = const.$ contours. The drain current is therefore continuous in V_{DS}. In V_{GSE}, the current again is forced to be continuous by the interpolation function used.

The partial derivative terms, for example g_{ds}, are in general discontinuous. A quadratic interpolation is used in V_{DS}, so g_{ds} has a step discontinuity about each data point *unless* the T_{DS} data are locally quadratic. An illustration appears in Figure 4.9. One way to solve the problem shown in the figure is by using a spline function along each V_{GSE} contour. But, if a $V_{GSE} = const.$ curve has for example 25 points along it, then a spline fit of the 25 points requires [18]:

$$4(25 - 1) = 96 \text{ coefficients for a cubic spline fit}$$

$$3(25 - 1) = 72 \text{ coefficients for a quadratic spline fit.}$$

Instead, if more points are stored, then the discontinuity in g_{ds} can usually be reduced to an insignificant level (within the error tolerances of the program) with fewer points than the number of coefficients required to spline-fit the original 25 points.

The reduction of the step discontinuity through adding points to T_{DS} is easily illustrated. Assume the points along a $V_{GSE} = const.$ curve are separated by $0.2V$ in V_{DS}. This corresponds to only 20 points total on the

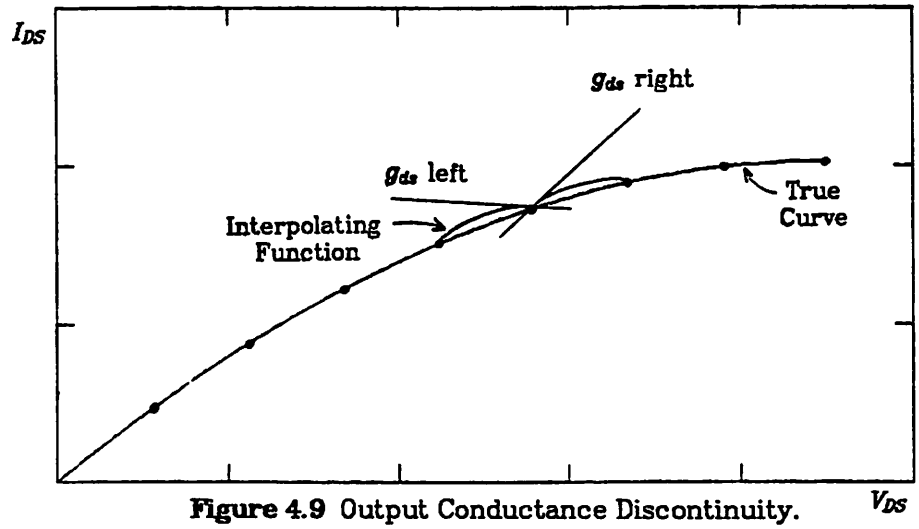

Figure 4.9 Output Conductance Discontinuity.

curve if $V_{DSAT} = 4V$. Assume also that the drain current varies as the cube of V_{DS}, viz.,

$$I_{DS} = 20 \cdot 10^{-6}\left[4.0 V_{DS} + 0.3 V_{DS}^2 - 0.1 V_{DS}^3\right]. \tag{4.1}$$

Equation (4.1) corresponds to Eq.(2.11) with $V_{GSE} = 4V$ and with $\lambda = 0.2$, which is an anomalously large value. It is thus fair to say that Eq.(4.1) represents a difficult test for reducing the g_{ds} discontinuity. The exact expression for g_{ds} results from differentiation of Eq.(4.1), i.e.,

$$g_{ds} = 20 \cdot 10^{-6}\left[4.0 + 0.6 V_{DS} - 0.3 V_{DS}^2\right]. \tag{4.2}$$

The magnitude of the step discontinuity in g_{ds} is examined at $V_{DS} = 3V$. The quadratic interpolation requires some neighboring points about $V_{DS} = 3V$. Using Eq.(4.1) gives

$$(V_{DS_1}, I_{DS_1}) = (2.80,\ 2.27136\text{E-}04)$$

$$(V_{DS_2}, I_{DS_2}) = (3.00,\ 2.40000\text{E-}04)$$

$$(V_{DS_3}, I_{DS_3}) = (3.20,\ 2.51904\text{E-}04).$$

Computing g_{ds} via the derivative of the quadratic interpolation function generates the results in Table 4.1.

Exact g_{ds}	Left g_{ds}	Right g_{ds}	Step in g_{ds}	Step Over Exact
6.200E-5	6.320E-5	6.080E-5	.240E-5	3.87%

Table 4.1

In Table 4.1, "Left" and "Right" refer to values from the segments to the left and right of $V_{DS} = 3V$, respectively, evaluated at $V_{DS} = 3V$. This is depicted in Figure 4.9.

CHAPTER 5

ONE-DIMENSIONAL EMPIRICAL MODEL

5.1. INTRODUCTION

The two-dimensional empirical model has a marked storage advantage over a simple three-dimensional approach. The most memory-intensive part of the 2-d model requires less than n^2 storage locations.

A further storage reduction is possible, however, if the information in $T_{DS}(V_{DS}, V_{GSE})$ can be decoupled so that only one-dimensional tables and/or functions are required. In addition, a 60% gain in computational efficiency can be achieved in practice by the decoupling.[1] This decoupling is implemented in the one-dimensional empirical model and constitutes its basic feature. The decoupling is accomplished by an origin-shifting transformation described below. The origin-shifting technique was originally derived and implemented by Newton [20] as part of a simple timing-analysis table look-up model which was subsequently extended to nonquadratic devices [14]. Further extensions of the model, primarily to make it applicable to circuit simulation, were contributed by this research. These extensions are the addition of interpolation, the addition of g_m and g_{mbs} calculations, a reformulation of the g_{ds} calculation, data set storage using spline functions, and a slight reformulation of the saturation-region portion of the model.

Historically, Newton's form of this model precedes the two-dimensional model by several years. The 1-d model is nevertheless presented after the

[1] This result is presented in detail in Chapter 6.

2-d model because it represents a further decoupling of the independent variables over the 2-d case.

5.2. MODEL DESCRIPTION

The major difference between the two-dimensional and one-dimensional empirical models lies in the storage of the linear-region $I-V$ characteristics. Both models remove the V_{BS} dependence, reducing the storage dimensionality from three to two, in the same manner. That is, V_{BS} is removed by defining

$$V_{GSE} = V_{GS} - S_1(V_{BS})$$

and thus resulting in $I_{DS} = I_{DS}(V_{DS}, V_{GSE})$. Both models store the data set for the saturation region in one-dimensional functions. However, the one-dimensional model effectively decouples V_{DS} and V_{GSE} so that functions of only a single independent variable can be used to store the linear-region $I-V$ characteristics. The transformation and original extensions of the model are included here for completeness.

5.2.1. Origin-Shifting Transformation The origin-shifting transformation is illustrated in this subsection through the use of the simple analytical MOSFET model described in Chapter 2. The extension of the model to non-quadratic devices is presented in the next subsection.

Equation 2.8 states that

$$I_{DS} = K\left[V_{GSE}V_{DS} - \frac{(V_{DS})^2}{2}\right], \qquad K = \frac{\mu_n C_{ox} W}{L} \tag{5.1}$$

in the linear region (i.e., $V_{DS} < V_{GSE}$). The family of curves which follows from Eq.(5.1) is shown in Figure 5.1 .

Now if the number of points on the V_{GSE} contour is doubled to 40, the interval in V_{DS} becomes $0.1V$ rather than $0.2V$. Repeating the calculations of the preceding paragraph results in the numbers in Table 4.2.

Exact g_{ds}	Left g_{ds}	Right g_{ds}	Step in g_{ds}	Step Over Exact
6.200E-5	6.260E-5	6.140E-5	.120E-5	1.94%

Table 4.2

Comparison of Tables 4.1 and 4.2 demonstrates that in this example the discontinuity is reduced in a linear manner by doubling the number of points stored to 40 points from the original 20.

Similar comments apply for g_m. The discontinuities arise at each $V_{GSE} = const.$ curve; they can be reduced by increasing the number of curves. The amount of discontinuity in g_{mbs} is fixed by g_m because the derivative of the S_1 spline is continuous, and

$$g_{mbs} = g_m \frac{-dS_1}{dV_{BS}} .$$

4.4.2. Out-of-Bounds Linear and Linear-Saturation Transition As already stated, current continuity is guaranteed by the two-dimensional model. The linear-to-saturation transition area does not have a continuity problem in g_{ds} if V_{GSE} is equal to one of the values on a stored contour, because of the saturation voltage definition.[4]

There are two places where g_{ds} can be discontinuous, both for intermediate V_{GSE} values. Figure 4.10 depicts the problem areas. A sufficient

[4]See appendix.

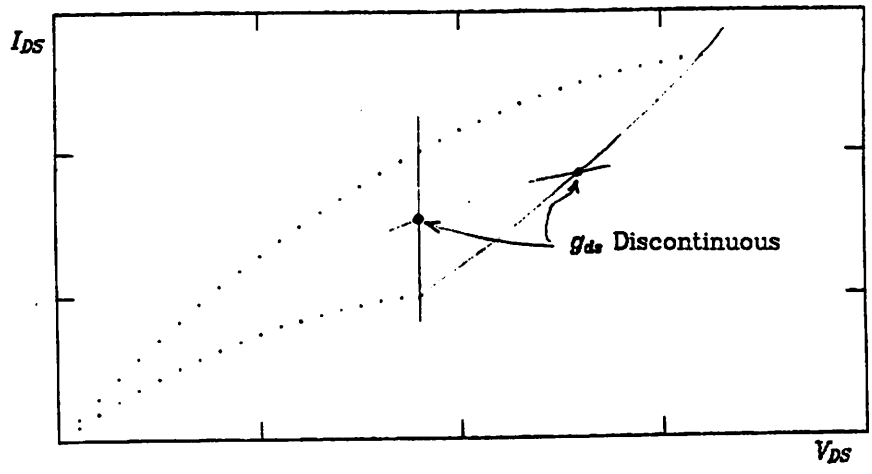

Figure 4.10 Transition and Out-of-Bounds Discontinuities.

number of $V_{GSE} = const.$ curves (on the order of 10 for a circuit with a maximum V_{GS} of 5 volts) must be stored in T_{DS} to make the discontinuity acceptably small. Note that g_m and g_{mbs} are continuous in the out-of-bounds area and across the transition region.

4.4.3. Saturation Region The spline functions which border the saturation region have continuous, smooth derivatives. Thus, all the partial derivatives are continuous in saturation.

Chapter 6 contains results on the fitting ability of the two-dimensional empirical model. Results on convergence and model evaluation time are also presented in Chapter 6.

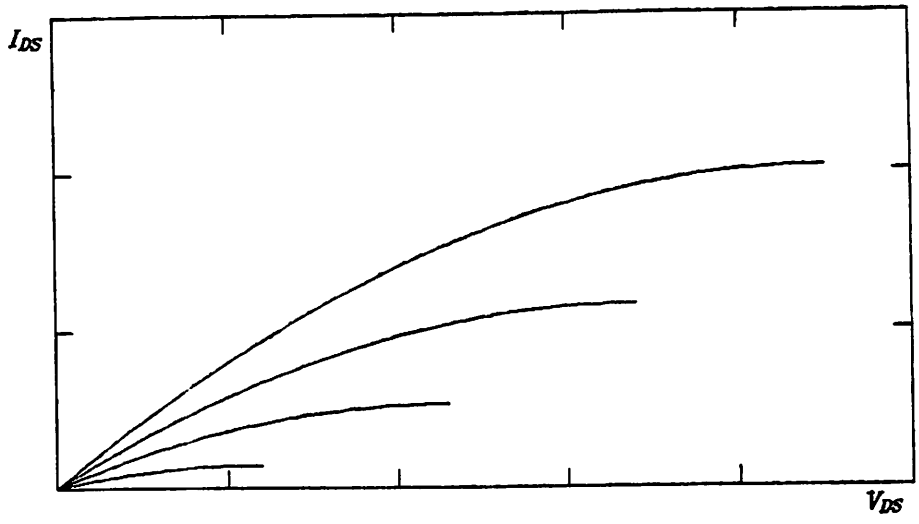

Figure 5.1 Output Characteristics From Eq.(5.1)

Examination of Figure 5.1 shows that each output characteristic curve seems to contain the one beneath it as its right-most segment, as shown in Figure 5.2. If this is the case, the single characteristic for the largest V_{GSE} value contains *all* of the linear-region $I-V$ information, provided it is *origin-shifted* appropriately. It is easy to show that an origin-shift does give an exact representation of the characteristics predicted by Eq.(5.1). To this end, assume that one curve is given, at $V_{GSE} = V_{GSE_{max}}$. The problem is to shift this maximum characteristic for $V_{GSE} < V_{GSE_{max}}$. If I_{DS} is written as

$$I_{DS} = K\left[V_{GSE_{max}}(V_{DS} + \Delta V) - \frac{(V_{DS} + \Delta V)^2}{2}\right] - K\left[V_{GSE_{max}}\Delta V - \frac{(\Delta V)^2}{2}\right] \tag{5.2a}$$

$$= K\left[V_{GSE_{max}}V_{DS} - \frac{V_{DS}^2}{2} - V_{DS}\Delta V\right], \tag{5.2b}$$

then the correct result is achieved for

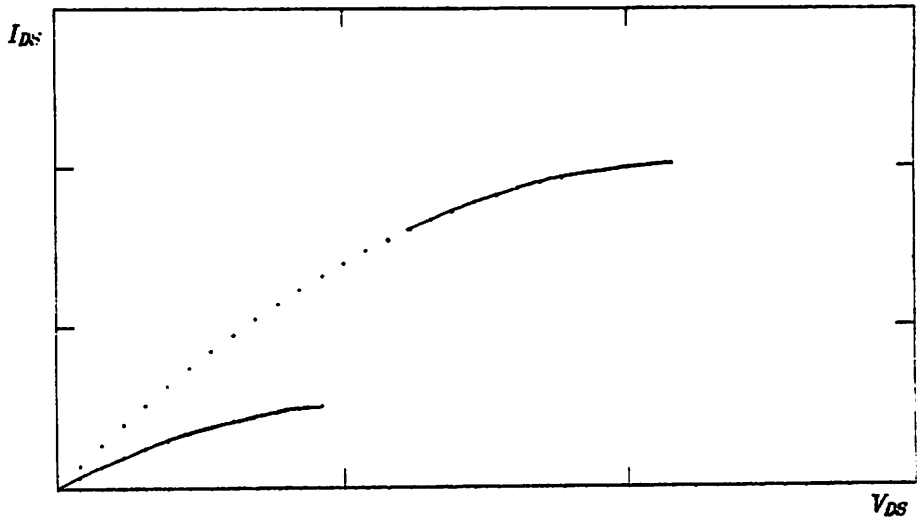

Figure 5.2 Similarity of Output Characteristics

$$\Delta V = V_{DSAT_{max}} - V_{DSAT} = V_{GSE_{max}} - V_{GSE}. \tag{5.3}$$

That is,

$$I_{DS} = K\left[V_{GSE_{max}}V_{DS} - \frac{V_{DS}^2}{2} - V_{DS}(V_{GSE_{max}} - V_{GSE})\right]$$

$$= K\left[V_{GSE}V_{DS} - \frac{V_{DS}^2}{2}\right]$$

as desired. Pictorially, the process of shifting the origin of the output characteristic for $V_{GSE} = V_{GSE_{max}}$ is shown in Figure 5.3 . The portion of the characteristic at the second set of axes has been transformed to model the dashed characteristic at $V_{GSE} < V_{GSE_{max}}$.

The transformation predicts the correct I_{DSAT} values (for this simple analytical model) if Eq.(5.2a) is rewritten as

$$I_{DS} = K\left[V_{GSE_{max}}[min(V_{DS}, V_{GSE}) + \Delta V] - \frac{[min(V_{DS}, V_{GSE}) + \Delta V]^2}{2}\right]$$

$$- K\left[V_{GSE_{max}}\Delta V - \frac{(\Delta V)^2}{2}\right]. \tag{5.4}$$

In the saturation region, $V_{DS} \geq V_{DSAT}$ $(= V_{GSE})$, and Eq.(5.4) becomes

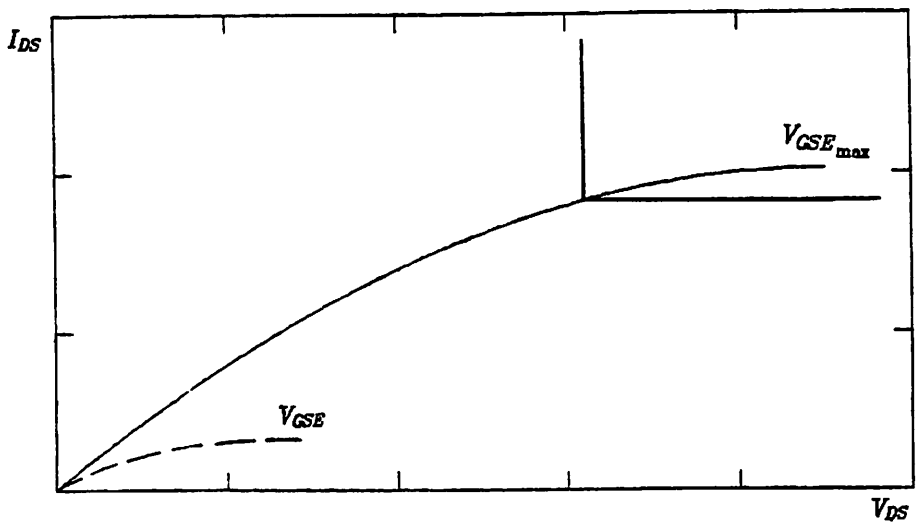

Figure 5.3 Origin-Shifting Transformation

$$I_{DS} = K\left[V_{GSE_{max}}(V_{GSE} + \Delta V) - \frac{(V_{GSE} + \Delta V)^2}{2}\right] - K\left[V_{GSE_{max}}\Delta V - \frac{(\Delta V)^2}{2}\right] \qquad (5.5)$$

$$= K\left[V_{GSE_{max}}V_{GSE} - \frac{(V_{GSE})^2}{2} - V_{GSE}(\Delta V)\right],$$

and since $\Delta V = V_{GSE_{max}} - V_{GSE}$,

$$I_{DSAT} = K\frac{(V_{GSE})^2}{2}.$$

The first term of Eq.(5.5) is

$$I_{DSAT_{max}} = I_{DS}(V_{DSAT_{max}}, V_{GSE_{max}}),$$

the current at the saturation point for the $V_{GSE} = V_{GSE_{max}}$ characteristic.

The origin-shifting transformation relies on the quadratic form of the $I_{DSAT}-V_{DSAT}$ function,

$$I_{DSAT} = \frac{K}{2}(V_{DSAT})^2.$$

The variation of I_{DSAT} with V_{DSAT} is not generally quadratic, though. The

extensions of the model which account for this and other departures from the simple theory are outlined in the next subsection.[2]

5.2.2. Extensions The original generalization of the 1-d model can be summarized as [14]

$$I_{DS} = T_{DS}(min(V_{DS}, V_{GSE}) + \Delta V) - T_{DS}(\Delta V) + T_G(V_{GSE}) \cdot V_{DS} \tag{5.6a}$$

$$\Delta V = T_S(V_{GSE_{max}}) - T_S(V_{GSE}) = V_{DSAT_{max}} - T_S(V_{GSE}) \tag{5.6b}$$

$$V_{GSE} = V_{GS} - T_B(V_{BS}) . \tag{5.6c}$$

The saturation voltage is allowed to vary with V_{GSE} through T_S, and the output conductance in saturation can change with V_{GSE} through T_G. This form of the model requires only four one-dimensional tables,

$$T_{DS}(V_{DS}) \tag{5.7a}$$

$$T_B(V_{BS}) \tag{5.7b}$$

$$T_S(V_{GSE}) \tag{5.7c}$$

$$T_G(V_{GSE}) . \tag{5.7d}$$

The 1-d model, with the generalizations of Eqs.(5.6), has been shown to fit small-geometry MOSFETs well [14].

5.2.3. Revised Model The model specified by Eqs.(5.6) was primarily intended for use in a type of timing simulator that does not require g_m, g_{mbs}, continuous current, or continuous g_{ds}, all of which are needed in circuit simulation. Equations (5.6) comprise the starting point for the one-dimensional model of this project. Some of the changes performed make the model useable for circuit simulation. Other alterations lower the requisite memory allocation by replacing look-up tables with cubic spline fitting func-

[2] An interesting fact to note is that the I_{DSAT}–V_{DSAT} characteristic is a rotated version of the linear-region $I_{DS}(V_{DS}, V_{GSE_{max}})$ characteristic for this simple model. This can be proven by a change-of-variable operation performed on Eq.(5.1) .

tions.

Four functions are used to store the data set as before. However, all of the functions are cubic splines, rather than tables. The functions which contain the data set for the revised model are:

$$I_{DS_{max}} = S_{DS}(V_{DS}) \tag{5.8a}$$

$$V_T = S_1(V_{BS}) \tag{5.8b}$$

$$V_{DSAT} = S_2(V_{GSE}) \tag{5.8c}$$

$$g_{ds_{sat}} = S_3(V_{GSE}) . \tag{5.8d}$$

Equations (5.8a), (5.8b), (5.8c), and (5.8d) are cubic splines that represent the same portions of the characteristics as T_{DS}, T_B, T_S, and T_G, respectively. The replacement of the tables with splines has resulted in a model with general fitting ability, as well as continuous drain current and partial derivatives. Also, the original model does not move all of the linear-region characteristics' V_{GSE} dependence to ΔV, because of the "$min(V_{DS}, V_{GSE})$" term in Eq.(5.6a). That minor restriction has been removed.

5.3. CURRENT AND CONDUCTANCE CALCULATIONS

The I_{DS} and g_{ds} calculations were done the same way for both the linear and the saturations regions in the original formulation of the 1-d model (Eqs.(5.6)). The revised model performs the current and the partial derivative calculations differently for each region, which results in a gain in computational efficiency. The initial step of determining whether the operating point lies in the linear region or the saturation region is the same as that of the 2-d model:

$$V_{GSE} = V_{GS} - S_1(V_{BS});$$

$$V_{DSAT} = S_2(V_{GSE});$$

if ($V_{DS} \geq V_{DSAT}$) then

transistor is in saturation;

else

transistor is in linear.

5.3.1. Normalization The first step in the generation of the data set for this model is to remove the effect of the saturation-region output conductance $g_{ds_{sat}}$ from the drain current characteristics. A set of output characteristics is the input to the data set generation programs. For each curve, corresponding to a particular V_{GSE}, the value for $g_{ds_{sat}}$ is found. Then, the current at each data point is normalized by subtracting off the term $(V_{DS})(g_{ds_{sat}})$. The effect of $g_{ds_{sat}}$ is added back into the drain current when the model is evaluated, such that continuity is maintained across the linear-to-saturation transition. The normalization simplifies the characterization procedure and results in faster model evaluation, as will be shown in the following subsection.[3]

5.3.2. Drain Current In the saturation region, the drain current is given by

$$I_{DS} = S_{DS}(V_{DSAT_{max}}) - S_{DS}(\Delta V) + S_3(V_{GSE})V_{DS} \tag{5.9a}$$

$$= I_{DSAT_{max}} - S_{DS}(\Delta V) + g_{ds_{sat}}V_{DS}\,, \tag{5.9b}$$

where $I_{DSAT_{max}}$ denotes the current at the saturation point for the $V_{GSE} = V_{GSE_{max}}$ characteristic and $\Delta V = V_{DSAT_{max}} - S_2(V_{GSE})$. If the normalization described in the previous subsection was not carried out, the current equation would instead be

[3]The data set generation procedure is outlined more fully in the appendix.

$$I_{DS} = S_{DS}(V_{DSAT_{max}}) - S_{DS}(\Delta V) - S_3(V_{GSE_{max}})V_{DSAT_{max}} + S_3(V_{GSE})V_{DS} ,$$

which is less efficient to compute.

A simple means is used to account for the non-zero saturation-region $g_{ds_{sat}}$ in this model. Spline S_3 contains $g_{ds_{sat}}$ as a function of V_{GSE} and is used to model the effect. The spline provides a term which is added to the basic model, as shown in Eqs.(5.9), such that the current in saturation has the form

$$I_{DS} = I_{DSAT} + g_{ds_{sat}}V_{DS}$$

where $I_{DSAT} = I_{DSAT_{max}} - S_{DS}(\Delta V)$. Thus, the output characteristics in saturation are approximated as linear in V_{DS} with their slopes dependent upon V_{GSE}. This formulation has the advantage that the partial derivative of I_{DS} with respect to V_{DS} is continuous across the linear-to-saturation transition, because the $S_3(V_{GSE})V_{DS}$ term is added to the current in the linear region also as shown next.

In the linear region, the drain current is given by

$$I_{DS} = S_{DS}(V_{DS} + \Delta V) - S_{DS}(\Delta V) + S_3(V_{GSE})V_{DS} \qquad (5.10a)$$

$$= S_{DS}(V_{DS} + \Delta V) - S_{DS}(\Delta V) + g_{ds_{sat}}V_{DS} \qquad (5.10b)$$

with ΔV defined as before. The drain current equation would be

$$I_{DS} = S_{DS}(V_{DS} + \Delta V) - S_{DS}(\Delta V) - S_3(V_{GSE_{max}})V_{DS} + S_3(V_{GSE})V_{DS}$$

without normalization of the data set.

The current calculation for the saturation region is simpler than that for the linear region. The partial derivative calculations can also be simplified when the operating region is known, as described in the next subsection.

5.3.3. Partial Derivatives The S_{DS} spline depends on a single independent variable, V_{DS}. Under the origin-shifting transformation, the difference between two values from S_{DS} is used in the calculation of the drain current. The two values are for two different *effective* V_{DS} values, $(V_{DS} + \Delta V)$ and (ΔV). From this point on, the following symbols are used:

$$V_{DSE_1} = V_{DS} + \Delta V$$

$$V_{DSE_2} = \Delta V .$$

By its definition $\Delta V = V_{DSAT_{max}} - S_2(V_{GSE})$ is a function of V_{GSE} alone. Thus,

$$\frac{\partial S_{DS}}{\partial V_{DS}} = \frac{dS_{DS}}{dV_{DSE}} = S'_{DS} \tag{5.11a}$$

$$\frac{\partial S_{DS}}{\partial V_{GSE}} = \frac{dS_{DS}}{dV_{DSE}} \frac{\partial V_{DSE}}{\partial V_{GSE}} = S'_{DS} \frac{-dS_2}{dV_{GSE}}. \tag{5.11b}$$

where V_{DSE} refers to the appropriate effective V_{DS}. The derivative of S_{DS} can be calculated explicitly.

In the linear region, the partial derivatives follow from differentiation of Eq.(5.10). Calculation of output conductance g_{ds} is straightforward. Equation (5.10b) has only two terms which vary with V_{DS}, the first term and the last term; g_{ds} is thus given by

$$g_{ds} = \frac{\partial I_{DS}}{\partial V_{DS}} = \frac{\partial S_{DS}(V_{DS} + \Delta V)}{\partial V_{DS}} + g_{ds_{sat}}$$

$$= S'_{DS}(V_{DSE_1}) + g_{ds_{sat}} . \tag{5.12}$$

The ΔV variable depends on V_{GSE}, and S_3 is a function of V_{GSE}. Therefore, all of the terms in Eq.(5.10a) contribute to g_m, and the expression for g_m is

$$g_m = \frac{\partial I_{DS}}{\partial V_{GSE}}$$

$$= \frac{\partial S_{DS}(V_{DS} + \Delta V)}{\partial V_{GSE}} - \frac{\partial S_{DS}(\Delta V)}{\partial V_{GSE}} + \frac{\partial S_3(V_{GSE})}{\partial_{GSE}} V_{DS}$$

$$= S'_{DS}(V_{DSE_1}) \frac{-dS_2}{dV_{GSE}} - S'_{DS}(V_{DSE_2}) \frac{-dS_2}{dV_{GSE}} + \frac{dS_3}{dV_{GSE}} V_{DS} . \qquad (5.13)$$

The backgate conductance g_{mbs} is calculated from g_m and the derivative of S_1 via the chain rule.[4]

The partial derivatives are analogously computed in the saturation region. The current equation in saturation is

$$I_{DS} = S_{DS}(V_{DSAT_{max}}) - S_{DS}(\Delta V) + S_3(V_{GSE})\, V_{DS} . \qquad (5.14)$$

The output conductance is explicitly stored in S_3, i.e.,

$$g_{ds_{sat}} = S_3(V_{GSE}) . \qquad (5.15)$$

The transconductance is slightly more complicated than the output conductance. The partial derivative of Eq.(5.14) with respect to V_{GS} is

$$g_m = -\frac{\partial S_{DS}(\Delta V)}{\partial V_{GSE}} + \frac{\partial S_3(V_{GSE})}{\partial V_{GSE}} V_{DS}$$

$$= -S'_{DS}(V_{DSE_2}) \frac{-dS_2}{dV_{GSE}} + \frac{dS_3}{dV_{GSE}} V_{DS} . \qquad (5.16)$$

The backgate conductance is calculated in the same manner as in the linear region.

Comparison of Eqs.(5.15) and (5.16) to Eqs.(5.12) and (5.13) and comparison of Eqs.(5.10) to Eqs.(5.9) demonstrates the savings in computational effort for the saturation region over the linear region which occurs from performing the calculations in an operating-region-dependent manner.

5.4. CONTINUITY CONSIDERATIONS

Like the 2-d model, the structure of the 1-d model guarantees drain current continuity. Unlike the 2-d model, the 1-d model also has all first partial derivatives continuous as a consequence of the use of splines to store the

[4] If not stated all derivatives are evaluated at the appropriate operating point.

entire data set.

In the next chapter results on the fitting ability, the time needed for model evaluation, and the convergence properties of the one-dimensional empirical model are presented.

CHAPTER 6

RESULTS

6.1. INTRODUCTION

This chapter contains results on the performance of the empirical models. The fitting ability of each model is demonstrated. The time required to evaluate the present models as well as a simplified version of each is given and compared to the model-equation-evaluation times for the SPICE2 analytical models. The effect on the convergence properties of SPICE2 due to the use of the models is described. Some conclusions are drawn regarding the models, and suggestions for future work are outlined.

6.2. ACCURACY

The accuracy of the empirical models in fitting MOSFET $I-V$ data is demonstrated in this section. In Chapter 3, it is stated that the data set for an empirical model can be developed from several sources. Accordingly, the data set for the 2-d model example below has been generated from the SPICE2 LEVEL-2 analytical model, with parameter values for a short-channel device. The data set for the 1-d model example has been developed from measured data from a short-channel MOSFET.

6.2.1. Two-Dimensional Model The data set for this test of the 2-d model has been generated from the SPICE2 LEVEL-2 model with parameter values appropriate for a contemporary 1.35μm channel-length NMOS device [23], in the manner outlined in Appendix 1.

Figure 6.1 shows the data which has been loaded into the T_{DS} table. Figure 6.2 shows the actual data as points, and the fit from the 2-d model as solid lines. The characteristic for the intermediate V_{GSE} value noted in Figure 6.2 demonstrates the adequacy of the interpolations in V_{GSE}, while each of the characteristics demonstrate the adequacy of the interpolations in V_{DS}.

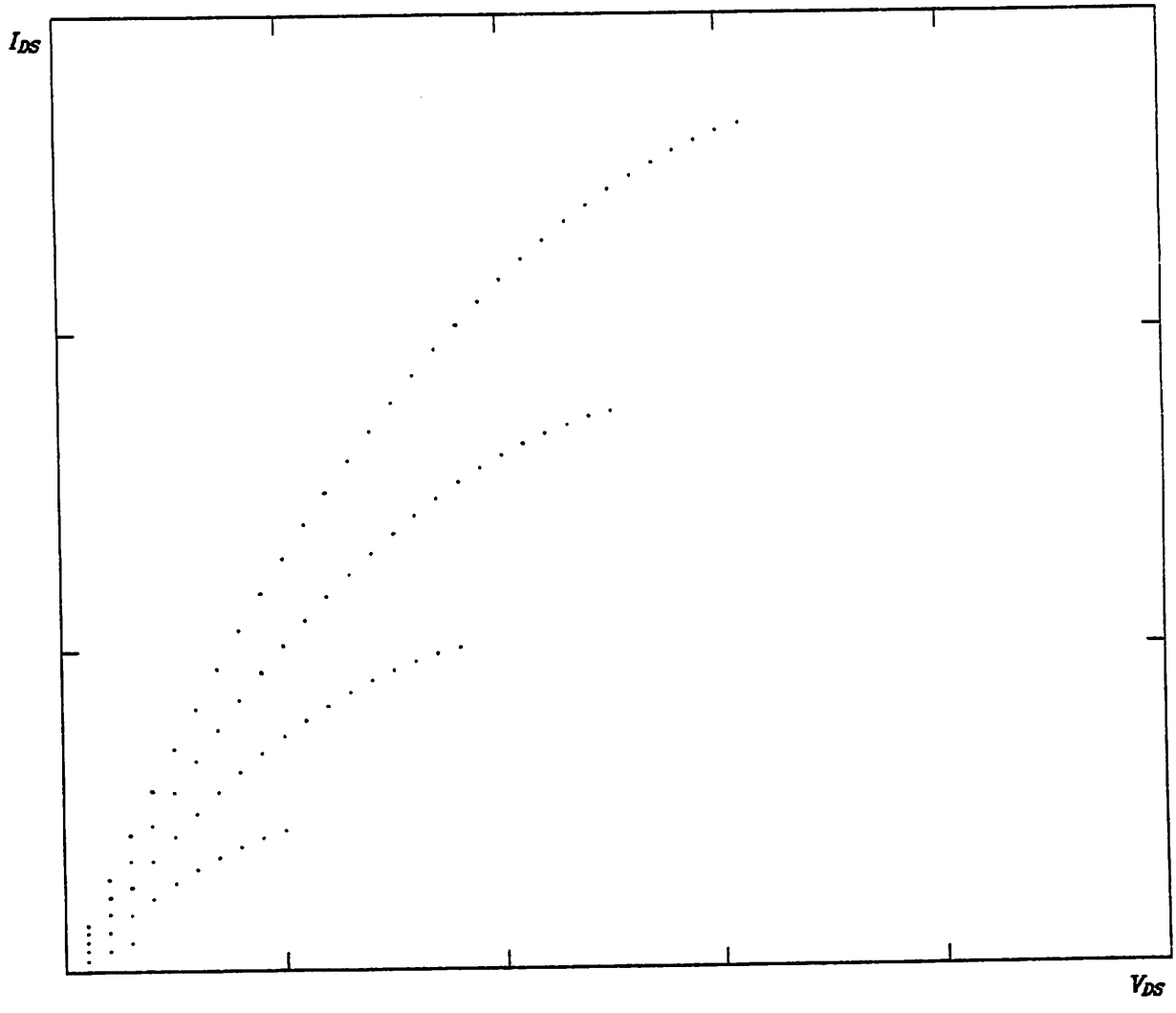

Figure 6.1 T_{DS} Data.

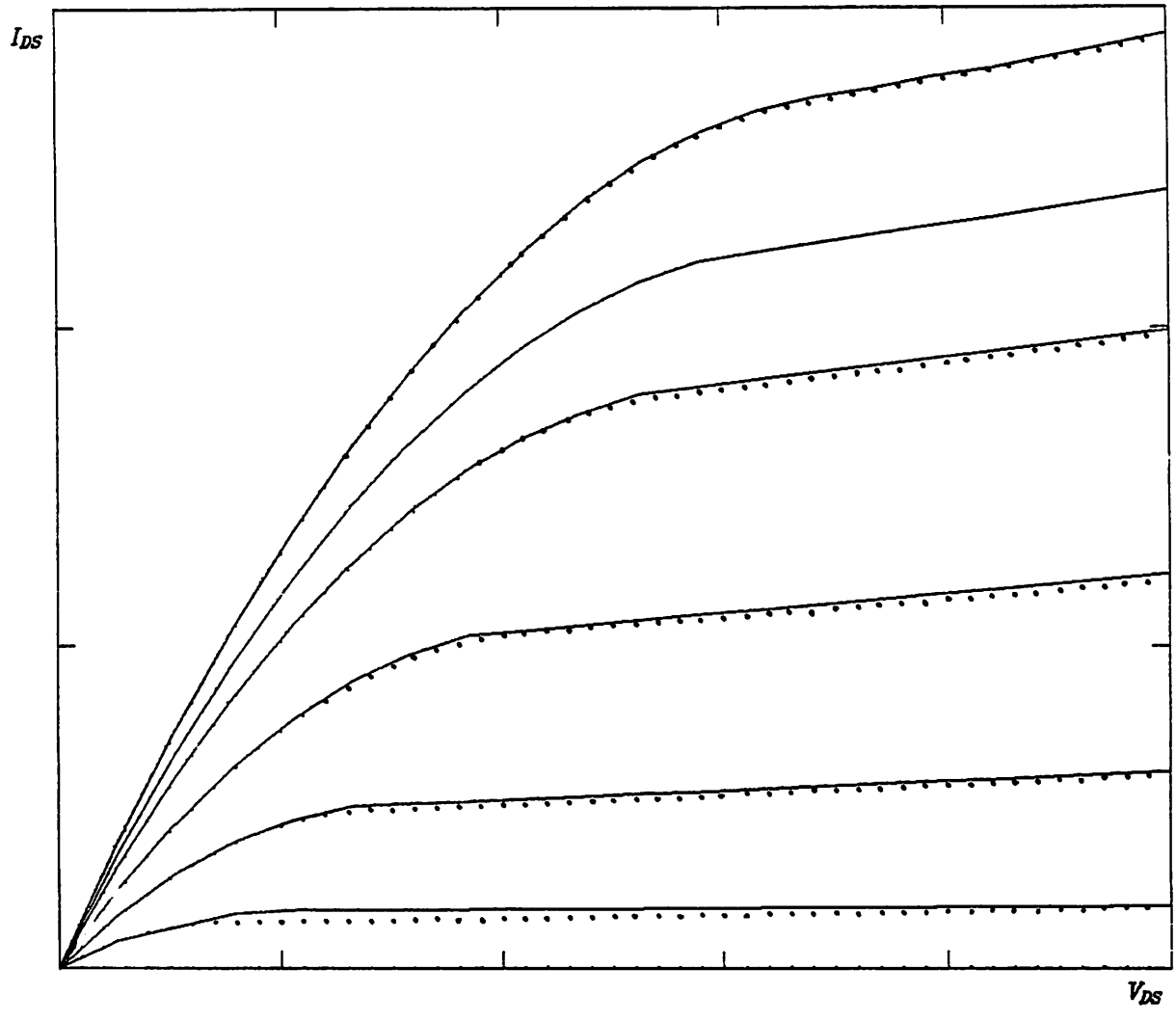

Figure 6.2 Output Characteristics from 2-d Model.

Linear-region output conductance curves from the two-dimensional model for this data set are shown in Figure 6.3. The curve corresponding to V_{GSE_4} is discontinuous at the boundary of the out-of-bounds linear region. The amount of discontinuity can be reduced by storing more data points in T_{DS}.

For this example, T_{DS} contains 91 points and each spline has been fit to 5 points. The total storage requirement is thus $91 + 16(5 - 1) = 155$ locations.

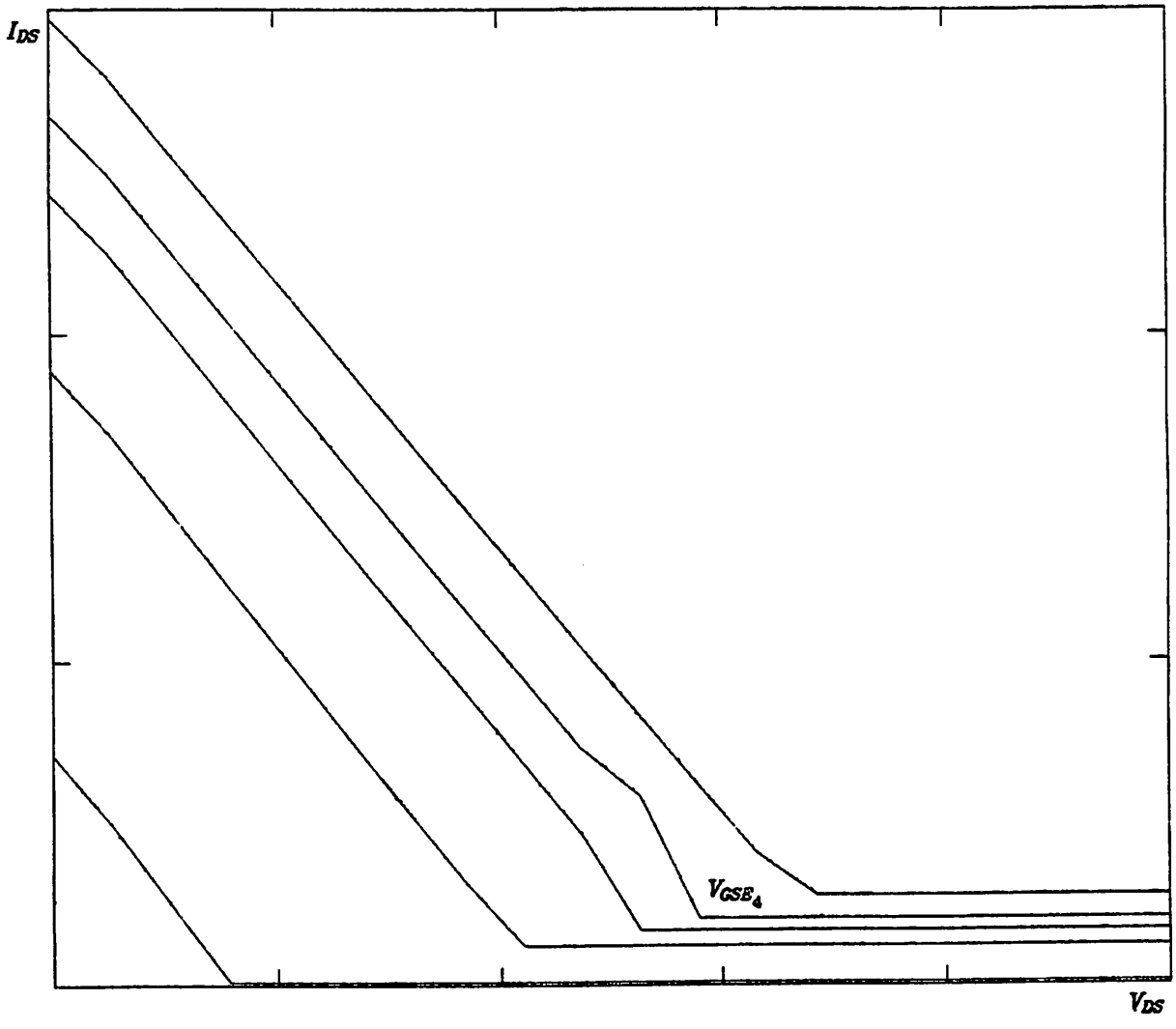

Figure 6.3 Linear-Region Output Conductance from the 2-d Model.

6.2.2. One-Dimensional Model The data set used for the accuracy test of the 1-d model presented in this subsection has been automatically developed using the programs described in the appendix and in [25]. The device used for the test is an n-channel depletion-mode transistor with drawn channel length and width of 2.5 μm and 50 μm, respectively.

Output characteristic curves from the 1-d model are depicted in Figure 6.4. The points represent measured data, while the solid curves are from the model. The curve for $V_{GSE_{max}}$ is the only one which is stored explicitly; the others are the result of the origin-shift operation. The fitting ability of the model for a difficult (i.e., short-channel, depletion-mode device) example is thus demonstrated.

The S_{DS} spline has been fit to eight data points in this example, which means that 28 spline coefficients are stored. Each of the other three splines is fit to five points, so the combined storage for their coefficients is 48 locations, resulting in a total storage requirement for the data set of only 76 locations.

6.3. SPEED OF EVALUATION

This section consists of results on the time required for computing the drain current and the partial derivatives via the two empirical models. Data is presented also for precursory versions of the models. Model equation evaluation times for SPICE2 analytical models are included for comparison.

6.3.1. Present Versions of the Empirical Models In Table 6.1 below, typical model evaluation times are given for each empirical model. The figures apply to a single evaluation of each model. The data in Table 6.1 does not include any charge calculations.

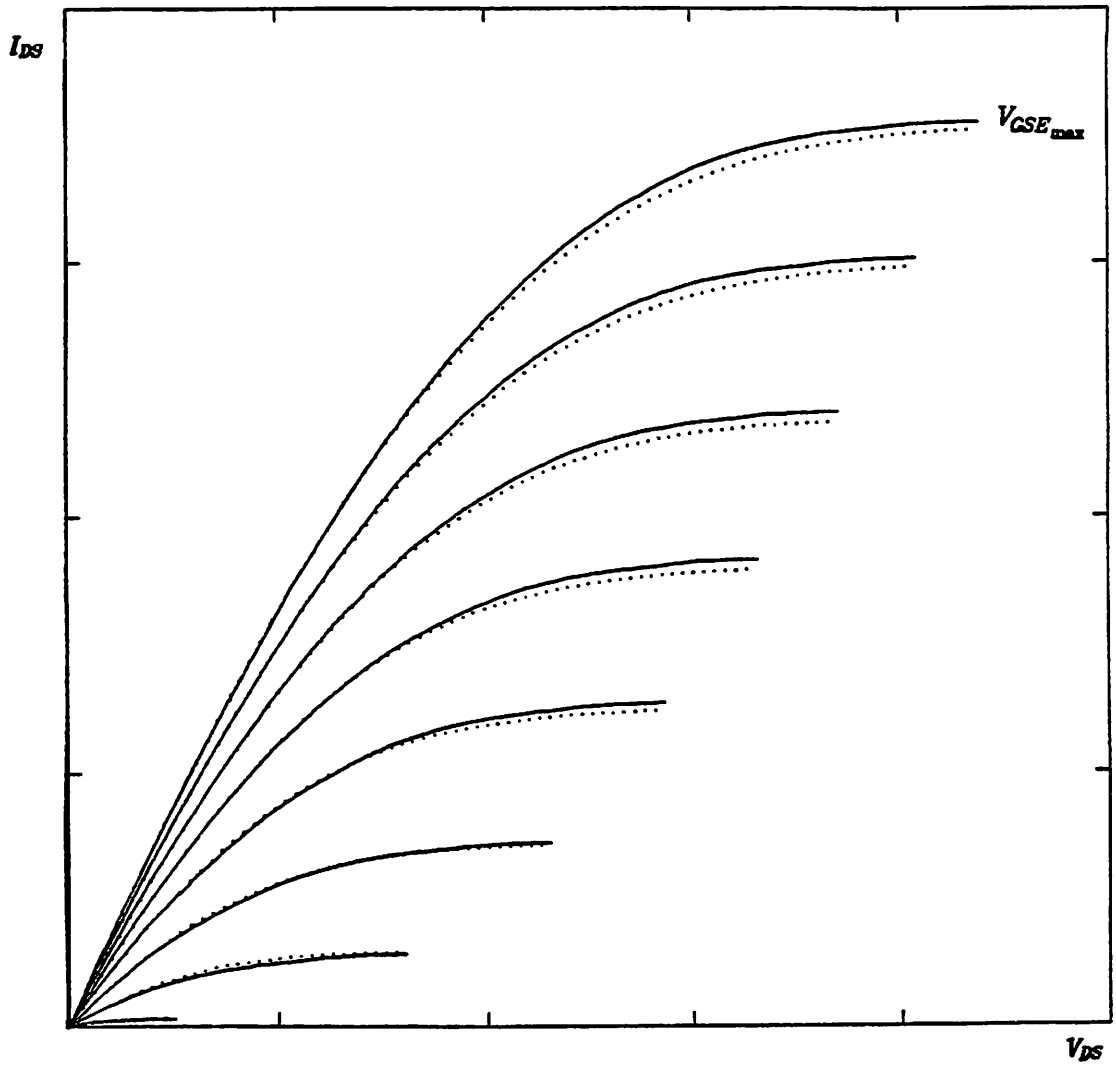

Figure 6.4 Output Characteristics from the 1-d Model.

Model	Typ. Eval. Time (ms)	Rel. Eval. Time
2-d	5.2	0.59
1-d	2.1	0.24
LEVEL-2	8.75	1.00

Table 6.1 Model Evaluation Times.

The column in Table 6.1 labeled "Relative Evaluation Time" demonstrates the speed advantage the empirical models have over an equivalent analytical model, i.e., the SPICE2 LEVEL-2 model. The 2-d model is nearly twice as fast as the LEVEL-2 model, and the 1-d model is over four times as fast as LEVEL-2. The SPICE2 LEVEL-3 model is up to 40% faster than LEVEL-2 [1]. Assuming that the LEVEL-3 model requires 5.25 ms for equation evaluation, the 2-d model is about equal to LEVEL-3 in efficiency, and the 1-d model is about 2.5 times more efficient than LEVEL-3.

6.3.2. Simplified Empirical Models A class of MOSFET circuits exists which can be simulated successfully without computing g_m and g_{mbs}. Included in this class of circuits are simple combinational switching circuits that do not have tightly-coupled feedback loops. Versions of the 1-d and 2-d empirical models appropriate for this type of circuit have been investigated in this research, and are reported in [13]. The data set is stored exclusively in tables in these simple models.

The results on model evaluation time from [13] for the simplified empirical models are given in Table 6.2.

Model	Typ. Eval. Time (ms)	Rel. Eval. Time
2-d	1.30	.15
1-d	1.07	.12
LEVEL-2	8.75	1.0
LEVEL-1	.53	.06

Table 6.2 Evaluation Times for Simplified Models.

Table 6.2 shows that the simple models both perform significantly faster in this test than the LEVEL-2 model.

The fact that the simple models do not calculate g_m and g_{mbs} does not affect the accuracy of the simulation, because circuit simulation programs converge to a solution. The absence of g_m and g_{mbs} can lead to nonconvergence for some circuits, though. When there is no convergence problem, the iteration count is typically increased by 10-20% over a similar simulation which incorporates g_m and g_{mbs}.

The evaluation speed of the 1-d model can be increased still further, if the accuracy constraint can be relaxed such that interpolation can be eliminated [20]. As reported in [20], the 1-d model at this degree of approximation is faster to evaluate than the LEVEL-1 model. However, the general utility of a model which is this approximate and which does not generate the transconductance and backgate conductance terms is limited.

6.3.3. SPICE2 Simulation Times The two empirical models have been tested in SPICE2 for dc and transient analyses. The models at their present state of development determine all of the dc quantities required at an operating point, but do not compute charge. Thus, in transient analysis the Meyer model [24] is used for the gate charge and the source and drain diffusions are represented by the familiar abrupt-junction capacitance

expression.

Total run times for SPICE2 are not greatly affected by the use of the empirical models. This is because the dc portion is a small part of the total model evaluation time for MOSFETs. Simulators that have MOSFET evaluation times which are dominated by the current and conductance calculations should have faster overall run times if empirical models are used.

For the circuits tested to date, the present versions of the empirical models do not significantly affect the convergence of SPICE2. However, more tests are needed to fully characterize the effects of the empirical models on the convergence properties of the program.

6.4. CONCLUSIONS

The empirical models which have resulted from this research fit the $I-V$ characteristics of modern MOSFETs well. The models are faster at generating the element values for the companion model than accurate analytical models. The empirical models have been successfully used in SPICE2. The data sets for the models are more readily generated than a parameter value set for an equivalent analytical model.

The 1-d model requires less storage than the 2-d model. But, which model provides the most fitting generality is not established. A wide variety of devices are currently being fit with each model to answer this question. The 2-d model decouples the $I-V$ information to a lesser degree than the 1-d model, and thus may be less approximate than the 1-d model.

For SPICE2, a speedup in the overall simulation time due to the speedup in the dc part of the model evaluation through the use of the empirical models is not apparent. Other types of simulators may see a change in the

overall simulation time. In particular, relaxation-based simulators which perform equation solution efficiently and are thus more affected by model evaluation time will see a bigger benefit when empirical models are used.

6.5. FUTURE WORK

There are several areas where more research could be carried out in empirical MOSFET modeling.

The effect on convergence resulting from the discontinuity of the partial derivatives of the 2-d model needs further investigation. The use of splines for the linear-region current table may provide a convenient means of eliminating this problem. Also, quadratic splines should be tested for the functions currently fit with cubic splines. Quadratic splines are faster to evaluate than cubic splines, and require fewer coefficients per number of points fit. However, if quadratic splines are used, more data points may be required to maintain accuracy.

Subthreshold conduction is not modeled. Subthreshold conduction is important in some present-day circuits, and will become more significant as device miniaturization continues [21]. An extension to the subthreshold region should be investigated in future work.

There is considerable disagreement on the proper analytical form for the MOSFET gate charge model. A study on the feasibility of representing the gate charge using empirical techniques should be performed.

Like the subthreshold-conduction effect, MOSFET punch-through becomes a greater problem as devices are scaled down [21]. Extension of the empirical models to the punch-through region is a potential research topic.

Perhaps the most interesting area for future work is in developing a hardware implementation of an empirical model. Since the data set storage requirements are modest, and since the arithmetic operations used in the model evaluation are simple, hardware implementation should be straightforward. This should yield very efficient model evaluation, and would be useful for example in a desktop computer which is limited in circuit simulation applications by floating-point performance.

APPENDIX 1

DATA SET GENERATION

Automatic data set generation programs for the 1-d and 2-d empirical models have been implemented by Hershbarger[25]. The methods used are outlined in this appendix for completeness.

Two-Dimensional Model

To begin the data set generation for the 2-d model, it is assumed that output characteristics have been measured or calculated as shown in Figure A.1. Each $V_{GSE} = const.$ curve provides values for T_{DS}, and each provides

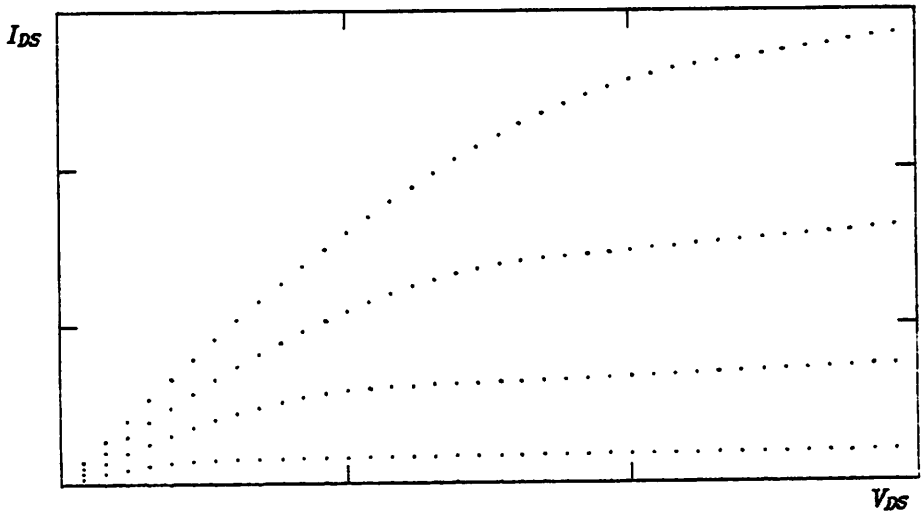

Figure A.1 Output Characteristics.

data points for the later determination of $V_{DSAT} = S_2(V_{GSE})$, $I_{DSAT} = S_3(V_{DSAT})$, and $I_{DMAX} = S_4(V_{GSE})$.

A smooth, differentiable fitting function, in this case a cubic spline, is fit to the first few points of a $V_{GSE} = const.$ curve, starting at $V_{DS} = 0$. The slope of the fitting function at the last point is used to extrapolate the characteristic to $V_{DS} = V_{DMAX}$. If too few points are fit, the resulting I_{DS} value at V_{DMAX} will be too large (Figure A.2). Another point is then added to those fit with the spline, and the extrapolation is repeated. When the extrapolated characteristic intersects the measured current at V_{DMAX} as shown in Figure A.3, the process is completed. If a data point does not exist at the saturation point, the suitable values for I_{DSAT} and V_{DSAT} can be generated from the spline fitting function in an iterative manner.

The points fit with the spline then are identified as the T_{DS} data for the V_{GSE} value of the characteristic. The value for V_{DSAT} is the V_{DS} value of the

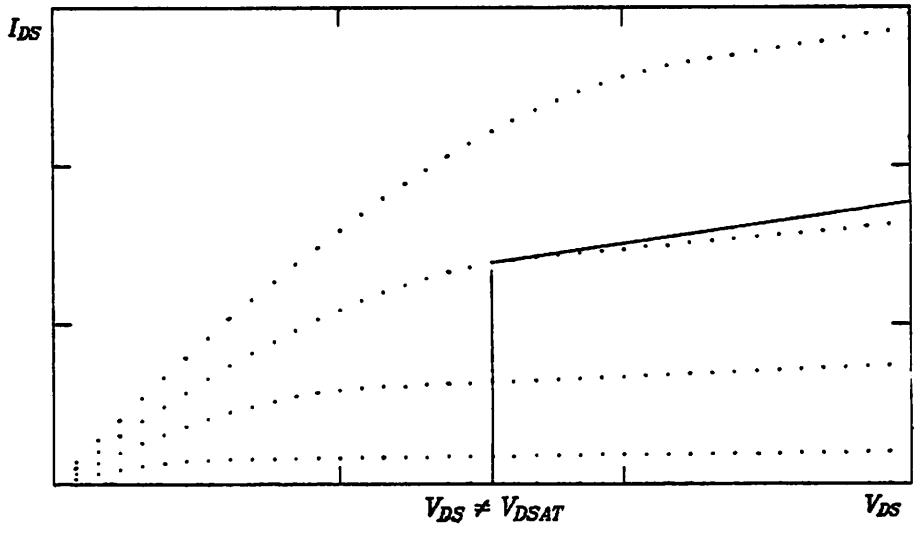

Figure A.2 Incorrect Extrapolation.

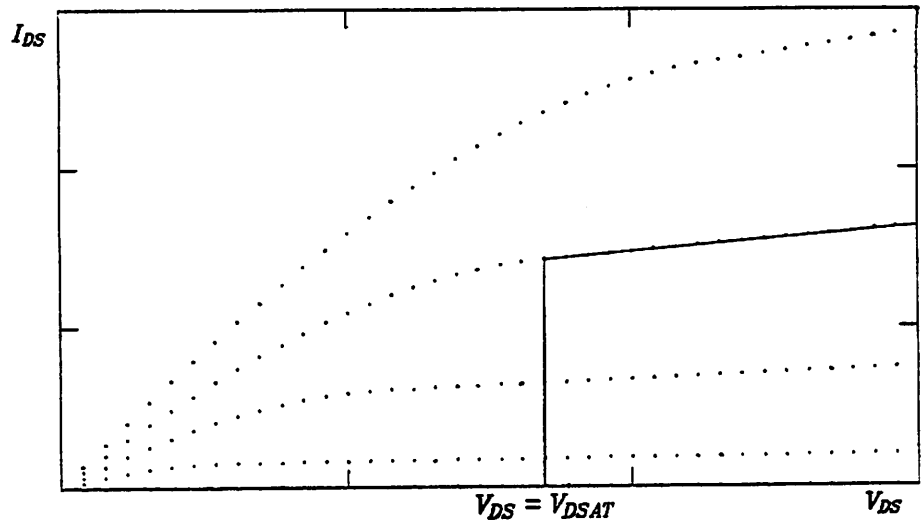

Figure A.3 Correct Extrapolation.

last point fit, and I_{DSAT} is the drain current at V_{DSAT}. The current I_{DMAX} at V_{DMAX} is already known. When the process is completed for all values of V_{GSE}, the data relating I_{DSAT} to V_{DSAT} and V_{DSAT} to V_{GSE} can be fit with cubic splines.

One-Dimensional Model

The data set generation program for the 1-d model takes as input the T_{DS} data found by the program for the 2-d model. The values for $g_{ds_{sat}}$ are also taken as input, and used to normalize the characteristics such that they have zero slope in saturation. This is accomplished by subtracting an amount $(V_{DS})(g_{ds_{sat}})$ from each T_{DS} drain current value.

The saturation voltage is adjusted for each curve so that I_{DSAT} is accurately predicted by the 1-d model. This step amounts to tuning the $V_{DSAT} = S_2(V_{GSE})$ function.

The final step in the data set generation is to tune the characteristic for $V_{GSE_{max}}$ such that the fractional error in I_{DS} predicted by the model is evenly distributed over the linear region. A weighted average of each characteristic determines the modifications made to the $V_{GSE_{max}}$ curve.

Threshold Voltage

The threshold voltage spline is calculated from the usual $V_T(V_{BS})$ data, obtained as described in [8] or [9]. In addition, an additive dependence on V_{DS} is accounted for, allowing V_T to be modeled as

$$V_T = S_1(V_{BS}) + S_5(V_{DS})$$

if desired. The S_5 spline is determined similarly to S_1.

REFERENCES

[1] S.Liu, "A Unified CAD Model for MOSFETs", ERL Memo No. UCB/ERL M81/31, University of California, Berkeley, May 1981.

[2] L.O.Chua and P-M.Lin, *Computer-Aided Analysis of Electronic Circuits*, Prentice-Hall, New Jersey, 1975.

[3] L.W.Nagel, "SPICE2 - A Computer Program to Simulate Semiconductor Circuits", ERL Memo No. ERL-M520, University of California, Berkeley, May 1975.

[4] A.Vladimirescu, "LSI Circuit Simulation on Vector Computers", ERL Memo No. UCB/ERL M82/75, University of California, Berkeley, October 1982.

[5] A.Vladimirescu, private communication.

[6] R.W.Hornbeck, *Numerical Methods*, Quantum Publishers, New York, 1975.

[7] C.F.Gerald, *Applied Numerical Analysis*, Addison-Wesley, Massachusetts, 1980.

[8] D.A.Hodges and H.G.Jackson, *Analysis and Design of Digital Integrated Circuits*, McGraw-Hill, New York, 1983.

[9] R.S.Muller and T.I.Kamins, *Device Electronice for Integrated Circuits*, Wiley, New York, 1977.

[10] H.Shichman and D.A.Hodges, "Modeling and Simulation of Insulated-Gate Field-Effect Transistor Switching Circuits", *IEEE J. Solid-State Circuits*,

Vol. SC-3, pp. 285-289, September 1968

[11] A.Vladimirescu, K. Zhang, A.R.Newton, D.O.Pederson and A. Sangiovanni-Vincentelli, "SPICE Version 2G User's Guide", University of California, Berkeley, 10 August 1981.

[12] A.Vladimirescu and S.Liu, "The Simulation of MOS Integrated Circuits Using SPICE2", ERL Memo No. UCB/ERL M80/7, University of California, Berkeley, October 1980.

[13] J.L.Burns, A.R.Newton and D.O.Pederson, "Active Device Table Look-Up Models for Circuit Simulation", *Proc. 1983 IEEE International Symposium on Circuits and Systems*, Newport Beach, California, pp. 250-253, May 1983.

[14] A.R.Newton, "Timing, Logic and Mixed-Mode Simulation for Large MOS Integrated Circuits", *Computer Design Aids for VLSI Circuits*, Sijthoff & Noordhoff International Publishers, The Hague, pp. 175-239, 1981

[15] T.Shima, T.Sugawara, S.Moriyama and H.Yamada, "Three-Dimensional Table Look-Up MOSFET Model for Precise Circuit Simulation" *IEEE J. Solid-State Circuits*, Vol. SC-17, pp. 449-454, June 1982.

[16] B.R.Chawla, H.K.Gummel, and P.Kozak, "MOTIS - An MOS Timing Simulator", *IEEE Trans. Circuits and Systems*, Vol. CAS-22, No. 12, pp. 901-909, December 1975.

[17] S.P.Fan, M.Y.Hsueh, A.R.Newton, and D.O.Pederson, "MOTIS-C: A New Circuit Simulator for MOS LSI Circuits", *Proc. 1983 IEEE International Symposium on Circuits and Systems*, Phoenix, Arizona, pp. 700-703, April 1977.

[18] C.deBoor, *A Practical Guide to Splines*, Springer-Verlag, 1978.

[19] J.Barby, J.Vlach and K.Singhal, "Polynomial Splines for FET Models", *Proc. 1983 IEEE International Symposium on Circuits and Systems*, Newport Beach, California, pp. 206-209, May 1983.

[20] A.R.Newton and D.O.Pederson, "Analysis Time, Accuracy and Memory Tradeoffs in SPICE2", *Conference Record*, Eleventh Annual Asilomar Conference on Circuits, Systems, and Computers, Asilomar, California, November 1977.

[21] S.M.Sze, *Physics of Semiconductor Devices, Second Edition*, Wiley, New York, 1981.

[22] H.C.Poon, "Vth and Beyond", Workshop on Device Modeling for VLSI, Burlingame, California, March 1979.

[23] I.Sakai, O.Kudoh, and H.Yamamoto, "High Speed 1 μm CMOS Technology", *IEDM Technical Digest*, San Francisco, California, December 1982

[24] J.E.Meyer, "MOS Models and Circuit Simulation", *RCA Review*, vol. 32, March 1971.

[25] R.A. Hershbarger, "Data-Set Generation for Empirical MOSFET Models", M.S. report, University of California, Berkeley, September 1983.