MEASUREMENT AND MODELING OF SMALL-GEOMETRY
MOS TRANSISTOR CAPACITANCES

by

JOHN JAMES PAULOS
S.B., Massachusetts Institute of Technology (1980)
M.S., Massachusetts Institute of Technology (1980)
E.E., Massachusetts Institute of Technology (1981)

SUBMITTED TO THE DEPARTMENT OF
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN ELECTRICAL ENGINEERING

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
September, 1984

(© Massachusetts Institute of Technology, 1984

Signature of Author ......................
Department of Electrical Engineering and Computer Science
August, 1984

Certified by .........................
Professor Dimitri A. Antoniadis
Thesis Supervisor

Accepted by ........................
Professor Arthur C. Smith
Chairman, Departmental Committee on Graduate Students
MEASUREMENT AND MODELING OF SMALL-GEOMETRY
MOS TRANSISTOR CAPACITANCES

by

John James Paulos

Submitted to the Department of Electrical Engineering and
Computer Science on August 15, 1984 in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Electrical Engineering

ABSTRACT

An integrated circuit technique for the measurement of
MOS device capacitances is described. This technique is
scalable and allows the measurement of gate (and sometimes
substrate) capacitances of minimum-geometry devices in any
MOS technology. Using this technique, gate and substrate
capacitances have been measured on devices as small as 4 \mu m
by 4 \mu m (1.1 \mu m by 1.5 \mu m effective) in a 5 \mu m CMOS process.
Data have been obtained for a variety of device geometries,
demonstrating short- and narrow-channel capacitance effects.
These data are used to evaluate simple models for these
effects which have been developed as part of this work.

In addition, the dynamic modeling of large-geometry
devices is reviewed. The two most significant quasi-static
capacitance models are those of Meyer and Ward. The relation-
ship between these two models is discussed, focusing on
the non-conservation of charge in circuit simulators based
on the Meyer model, and the partitioning of the channel
charge in the Ward model. Also presented is a non-quasi-
static analysis of the MOSFET which provides some insight
into the source of the nonreciprocal capacitances in the
Ward model. The small-signal admittances of the Meyer and
Ward models are compared to those of the non-quasi-static
analysis, highlighting the limitations of these quasi-static
models at high frequencies.

Thesis Supervisor: Dr. Dimitri A. Antoniadis

Title: Associate Professor of Electrical Engineering
# TABLE OF CONTENTS

Abstract .............................................................................................................. 3

Table of Contents ............................................................................................... 5

Acknowledgments ............................................................................................... 7

INTRODUCTION ................................................................................................. 9

CHAPTER 1: Fundamental Issues in Capacitance Modeling for Circuit Simulation .............................................. 17

1-1. Modeling electronic devices ................................................................. 18

1-2. Nonlinear multi-terminal capacitors ................................................... 24

CHAPTER 2: History of MOSFET Dynamic Modeling ................................................. 45

2-1. Static analysis of the MOSFET ......................................................... 46

2-2. The Meyer quasi-static capacitance model ........................................... 56

2-3. The Ward model and other charge-partitioning models ......................... 66

2-4. Non-quasi-static modeling of the MOSFET ......................................... 83

CHAPTER 3: Measurement of Small-Geometry MOS Transistor Capacitances ................................................ 105

3-1. Motivation for an on-chip capacitance measurement approach ................. 106

3-2. A summary of general capacitance measurements techniques ............... 112

3-3. The closed-loop coulombmeter technique ............................................. 126
CHAPTER 4: Modeling of Small-Geometry MOS Transistor Capacitances .................................. 155

4-1. Narrow-channel capacitance effects .............. 156

4-2. Short-channel capacitance effects ............. 197

CONCLUSIONS ................................................................. 229

Appendix A: Gate and Bulk Capacitance Data ................. 235

References ................................................................. 297
ACKNOWLEDGMENTS

I would like to thank Professor Dimitri Antoniadis for supervising this work, and Professors James Roberge, Paul Penfield, and John Wyatt for acting as readers. Each member of the thesis committee has made important contributions to my understanding of this subject, and their support is greatly appreciated. I would also like to thank my office mates, Michael Coln and Andrew Robinson, for their contributions during our many discussions of this and other topics.

This work has been supported in part by an IBM Fellowship. The CMOS processing was provided by the GTE Microelectronics group of Needham, Massachusetts, and I would particularly like to thank Jim Sims for funding the project and Michael DeSmith for his close cooperation and technical support.

Finally, I would like to thank my wife for her patience, and for her confidence in the ultimate success and completion of this work.
INTRODUCTION

In the 24 years since the invention of the Metal-Oxide-Semiconductor Field-Effect Transistor (MOSFET) [0.1], MOS integrated circuit technology has matured through a steady shrinking of device geometries and circuit interconnections. In the process, MOS technology has been proven to be superior to bipolar technology for high-density VLSI (Very Large Scale Integration) circuits. The recent announcements of 1 Mb memories [0.2,0.3] and a 450,000 transistor microprocessor [0.4] demonstrate the VLSI capabilities of MOS technologies which use small-geometry transistors with gate lengths of 1-1.5 μm.

Most of the reduction in device geometries that has been achieved to date has come through improvements in fabrication technology. The early contact pattern exposure (lithography) systems using visible light were replaced by systems using ultraviolet light, and these systems are increasingly being replaced by stepped reduction projection systems. For the future, electron beam and X-ray exposure systems are being explored which could provide resolution of better than 0.1 μm. Similarly, the wet chemical etching techniques of early MOS technologies are giving way to a variety of 'dry' patterning techniques, using reactive ion plasmas, which are capable of much finer line width defini-
tion. However, as lithography and definition techniques continue to improve, MOS technologies are increasingly limited by junction depths and lateral oxide encroachment, and the effects of these physical structures on device behavior.

Junction depths are determined by the movement of the ion-implanted source and drain impurities during the subsequent thermal annealing of the damage caused by the ion implantation. For modern MOS technologies, this implantation/annealing constraint causes junction depths to be on the order of 0.25–0.5 μm, depending on the impurity species. This effectively limits the minimum channel length, since short-channel effects arise when the channel length is reduced to the point where it approaches the junction depth.

Oxide encroachment refers to the tapered oxide region formed during the growth of the thick field oxide. This oxide taper reduces the effective device width by increasing the gate oxide thickness near the edges of the channel. The lateral extent of the encroachment is on the order of the field oxide thickness, typically around 0.5 μm. This limits the minimum device width, since narrow-channel effects arise when the channel width is reduced to the point where it is on the order of the lateral oxide encroachment.

The minimum device geometries are therefore constrained by the process technology and the degree to which short- and
narrow-channel effects can accurately be modeled and controlled. Not surprisingly, the modeling of these and other small-geometry effects has been the principal area of activity in MOS device research, at least since the early 1970's. Short-channel effects have been studied extensively since 1968 [0.5], and narrow-channel effects have been studied since 1973 [0.6]. Much of this work is limited to the study of threshold voltage \( V_T \) variations with channel length and width, and almost all of this work relates only to static (DC) behavior. To date, only one paper [0.7] which discusses the effects of short channels on the MOSFET device capacitances has appeared, and a similar treatment for narrow-channel effects has yet to appear.

This neglect of short- and narrow-channel capacitance modeling in the literature mirrors the history of capacitance modeling of large-geometry devices. The period from 1964, when the first careful static analysis of the MOSFET was presented [0.8], to 1970 saw the greatest activity in the dynamic modeling of the MOSFET. During this period, several low-frequency (quasi-static) capacitance models and high-frequency (non-quasi-static) equivalent circuit models were introduced in the literature. This activity was reviewed, in a 1970 book by Cobbold [0.9] which included a complete, low-frequency, incremental capacitance model. This model represented the state of the art in MOSFET dynamic modeling until 1978. Moreover, a similar model,
formalized by Meyer in 1971 [0.10], was almost universally used in digital computer circuit simulators such as SPICE [0.11] until very recently.

In 1978 however, Ward identified a fundamental problem with the Meyer model which causes non-conservation of charge for circuit simulations [0.12]. Ward proposed a new model which, in addition to solving the conservation of charge problem, predicted that the small-signal capacitances of the MOSFET should be nonreciprocal. Nonreciprocal capacitances were experimentally observed and reported, apparently for the first time, by Ward in [0.13]. One is left to conclude that a fundamentally incorrect model was widely used for over ten years because a careful experimental characterization of the MOSFET device capacitances was never undertaken to verify the model.

The principal goal of this work is to preclude a similar decoupling of theory and experiment in the modeling of small-geometry capacitance effects. (It is interesting to note that the work on the modeling of short-channel device capacitances [0.7] referred to above does not include experimental data to verify the model.) The problem in this case is that capacitance measurements for small-geometry devices are very difficult. For example, the total gate capacitance of a 1 μm by 1 μm device with 350 Å gate oxide is about 1 fF. For a high resolution measurement of this capaci-
tance, an instrument with resolution of 1 aF (atto Farad) is required. This is 100 times the resolution generally available from conventional capacitance instruments.

To solve this problem, an integrated circuit technique for the measurement of MOS device capacitances has been developed. This scalable technique is capable of measuring gate (and sometimes substrate) capacitances of minimum-geometry devices in any MOS technology. By using on-chip circuits, this technique avoids the limitations of conventional capacitance meters, and eliminates the interference produced by the stray capacitances associated with bonding or wafer probing. Using this technique, gate capacitances have been measured on devices as small as 4 μm by 4 μm in a 5 μm CMOS process. (Although these devices are large compared to those of other MOS technologies, the capability of measuring minimum-geometry devices should hold for 1 μm geometries in a 1 μm process and beyond.) Data have been obtained from a variety of device geometries which demonstrate short- and narrow-channel capacitance effects. Also, simple models have been developed for these short- and narrow-channel capacitances.

This document presents the new technique and models in conjunction with some background on large-geometry capacitance modeling. It also presents a non-quasi-static analy-
sis for a simplified MOSFET which gives some new insight into the nonreciprocal capacitances of the Ward model.

Chapter 1 presents the background material on dynamic modeling and multi-terminal capacitances which will be basic to the discussions in the following chapters. The formulation of models for circuit simulators, which motivates the use of multi-terminal notation rather than multi-port notation, is discussed. This chapter also explores the concept of a generalized, nonlinear multi-terminal capacitor, which provides a mathematical basis for the discussion of multi-terminal capacitance modeling, including nonreciprocal capacitances.

Chapter 2 reviews the history of dynamic modeling of the MOSFET. The Meyer and Ward models are described in detail, with a focus on the conservation of charge problem in the Meyer model and the partitioning of the channel charge in the Ward model. Several other recent models which propose alternative partitioning algorithms are also examined. A review of high-frequency modeling is presented, and a non-quasi-static analysis is performed which verifies the Ward model and provides some insight into the physical mechanisms behind the nonreciprocal capacitance terms.

Chapter 3 presents the on-chip capacitance measurement technique. General methods for capacitance measurement are reviewed highlighting the limitations and advantages of each
of these methods. A closed-loop measurement technique has been chosen and the implementation of this technique in both NMOS and CMOS technologies is discussed.

Chapter 4 discusses the modeling of short- and narrow-channel capacitances, drawing on data taken using a CMOS implementation of the on-chip measurement technique. A complete set of gate and bulk capacitances for several device geometries is included in Appendix A for future use.
CHAPTER 1

Fundamental Issues in Capacitance Modeling for Circuit Simulation

This chapter reviews some of the fundamental concepts associated with the dynamic modeling of semiconductor devices. First, an overview of device modeling is presented which highlights the wide variety of potential models and modeling applications. The application of chief concern for this work is large-signal circuit analysis using a digital computer circuit simulator. This application requires a model with a specific mathematical formulation, which leads to the indefinite multi-terminal notation that is used throughout this work.

Second, the concept of a general nonlinear multi-terminal capacitor is examined. This provides a framework for the later discussions of charge partitioning and non-reciprocal capacitance associated with the Ward model [1.1,1.2]. This section also introduces the violation of conservation of energy in nonreciprocal capacitors, which will be referred to in Chapter 2.
1.1 Modeling Electronic Devices

Modeling electronic devices is a two-step process. The physics of the device behavior must first be studied and understood before a model can be created to suit a particular kind of circuit analysis. Although these two steps may be closely coupled, they are nevertheless distinct and separate; for while there may be only one physical analysis that completely describes the device behavior, there may be many models for the device which are each useful for a different application. However, it is important that models remain closely tied to the underlying physics of the device; otherwise, the relationships among the model parameters and between the model parameters and the physical device geometry will be lost. Finally, since every model is created for a different application, it is important that the bounds on a model's applicability are clear.

1.1.1 Types of models and their applications

A wide variety of device models is generally needed to satisfy different applications. These models will vary according to the kind of circuit being analyzed, the type of analysis being performed, and the kinds of signals that are of interest. For example, for analog circuits a model which is as precise as possible may be desired, while for digital circuits a much simpler model may be used. Similarly, a
A fairly complete dynamic model is needed for the transient (timing) analysis of digital circuits, but for the logical analysis of digital networks, the MOSFET can be modeled as an ideal switch. Three characteristics of device models are of interest here: the form of the model, the class of input signals, and the frequency range of the model.

Device models are generally either complex mathematical models intended for computer-based calculations or simple mathematical or equivalent circuit models intended for rough hand analysis. Most design methodologies consist of an iterative, two-step process. The designer first uses simple device models and standard practice 'rules of thumb' to choose the circuit topology and the device geometries. Then the designer uses a digital computer to obtain an accurate simulation of the performance of the circuit. If the simulation indicates that the circuit does not meet the performance specifications, the process is repeated. The device model used in the computer simulation needs to be sufficiently accurate to insure that the fabricated circuit will meet the design specifications. On the other hand, the device model used in hand calculations needs only to be accurate enough to lead the designer in the right direction so that the design process will converge.

A device model for computer-based applications will almost always be mathematical, but a device model for hand
calculations may be either a mathematical model or a circuit equivalent model. The charge control technique for large-signal, bipolar switching calculations is an example of a mathematical hand calculation model, while the incremental, hybrid-pi model is an example of an equivalent circuit model for hand analysis. It is frequently the case that high-frequency, incremental models are equivalent circuit models. These models have the advantage that they make use of the designer's intuition for linear networks. In this study, we will be principally interested in models for computer simulation, where accuracy is particularly important. The specific mathematical form which these models must take is discussed in section 1.1.2.

The second distinction is that of small-signal versus large-signal models. Large-signal models apply over the full range of terminal voltages and are used for transient analyses, particularly of digital circuits. Small-signal models are applicable only around an operating point, and are generally used for AC (small-signal sinusoidal) analyses of analog circuits. Small-signal models can almost always be generated from large-signal models by incrementalization around an operating point; however, the converse is almost never true.

The third point of interest is the range of signal frequencies for which the model is designed. Device models
are generally characterized as static (DC), low-frequency, or high-frequency models. A static model describes the (static) terminal currents in terms of the applied terminal voltages. This is equivalent to a nonlinear multi-terminal resistor model of the device. Low-frequency models add capacitances which are based on quasi-static calculations of the device charges. A quasi-static charge description is one which expresses the instantaneous charges as functions of the instantaneous terminal voltages, i.e., there are no dynamics in the charge-voltage relationship. High-frequency models describe higher order frequency effects, and are generally based on linearizations of the underlying semiconductor transport and continuity equations. These models are frequently expressed as circuit equivalent models although this is not necessarily the case. Both low-frequency and high-frequency models for the MOSFET will be discussed in Chapter 2.

1.1.2 Models for circuit simulators

Most circuit simulators, such as SPICE [1.3], are based on nodal analysis or modified nodal analysis. In nodal analysis the network variables are the node voltages and the branch currents which enter (or leave) the nodes. For each element in the network, a device description which relates the device terminal currents to the terminal voltages is required. These I-V relationships are used to eliminate the
branch currents by applying Kirchoff's Current Law (KCL) at each node. This produces a set of simultaneous equations, \( \mathbf{f} \), of the form

\[
\mathbf{f}(\mathbf{V}) = 0
\]  \hspace{1cm} (1.1)

where \( \mathbf{V} \) is the vector of node voltages. If \( \mathbf{f}(\mathbf{V}) \) is of the form \( \mathbf{A}\mathbf{V} + \mathbf{B} \), Eq. 1.1 can be solved directly for the node voltages. (Given the node voltages, the branch currents can be calculated from the the element I-V relationships.) If this is not the case, an equation of the desired form can be achieved by linearizing \( \mathbf{f}(\mathbf{V}) \) around an initial guess for the node voltages, \( \mathbf{V}_0 \), producing

\[
\mathbf{f}(\mathbf{V}) \approx \mathbf{Y} \delta \mathbf{V}_0 + \mathbf{f}(\mathbf{V}_0) = 0
\]  \hspace{1cm} (1.2)

where \( \mathbf{Y} \) is the linearized admittance matrix. Solving (1.2) for \( \delta \mathbf{V}_0 \) yields an improved estimate of the node voltages, \( \mathbf{V}_1 \), where \( \mathbf{V}_1 = \mathbf{V}_0 + \delta \mathbf{V}_0 \). This can be used in a new linearization of \( \mathbf{f}(\mathbf{V}) \), as in (1.2), to find an even better estimate of \( \mathbf{V} \). This process, known as Newton-Raphson iteration, will generally converge as long as the initial guess for \( \mathbf{V} \) is close enough.

The advantage of nodal analysis is the ease with which the matrix \( \mathbf{Y} \) can be generated. In particular, the I-V relationship of each element can be linearized individually to form a submatrix \( \mathbf{Y}_i \). These submatrices are then summed to create the total matrix \( \mathbf{Y} \). For nonlinear transient
analysis, the element I-V relationships will include a
discretization in time of the dynamic behavior of the
device. This is discussed in Section 2.2.

The nodal analysis used by computer simulators requires
device models which can be described by terminal current-
voltage relationships. This is in contrast to state vari-
able methods, for which a state variable description is
used. Furthermore, an indefinite multi-terminal description
of the device is desired, rather than a multi-port descrip-
tion, to facilitate the construction of the linearized
admittance matrix. An indefinite N-terminal description is
equivalent to an N-port description of the same network plus
an arbitrary reference node, where each of the N port pairs
consists of one of the device terminals plus the reference
node. For example, an indefinite three-terminal description
of the drain current of the MOSFET would be expressed in
terms of \( V_G \), \( V_S \) and \( V_D \), rather than \( V_{GS} \) and \( V_{DS} \), which are
the two-port variables. This indefinite nodal notation is
used in the following discussion of multi-terminal capaci-
tors, and throughout the rest of this document.
1.2 Nonlinear Multi-Terminal Capacitors

The concept of a general nonlinear multi-terminal capacitor, which is fundamental to the discussion of dynamic modeling of the MOSFET, is developed in this section. First, the nonlinear two-terminal (one-port) resistor is reviewed as a point of comparison. Next, the nonlinear two-terminal capacitor is described, focusing on the fundamental role of charge and the distinction between large-signal and small-signal descriptions of nonlinear capacitors. Then the nonlinear two-terminal capacitor is generalized to \( N \) terminals, which leads to the definition of the incremental capacitances for the multi-terminal device and to the definition of the Indefinite Admittance Matrix (IAM). A discussion of nonreciprocity and the implication of non-conservation of energy in nonreciprocal capacitors follows. Finally, the experimental characterization of multi-terminal capacitors is described with some comments on the distinction between models and the measurement of real devices.

1.2.1 The two-terminal nonlinear resistor

A simple (two-terminal) nonlinear resistor (Fig. 1.1) is described by its constitutive relationship between branch current \( (i) \) and branch voltage \( (v) \). In general, this relationship is described by a path (or paths) of allowed points \( (i,v) \) in the current-voltage plane. If the device current
Figure 1.1 - The nonlinear two-terminal resistor.

can be expressed as a single-valued function of the device voltage, as in the example of a tunnel diode (Fig. 1.2a), the resistor is said to be voltage-controlled. If the device voltage can be expressed as a single-valued function of the device current, as in the example of a neon bulb (Fig. 1.2b), then the resistor is said to be current-controlled. In many cases, the constitutive relationship is strictly monotonic and the device is both voltage-controlled and current-controlled. This is the case for a non-zero linear resistor, shown in Fig. 1.2c.

The constitutive current-voltage relationship is a large-signal description of the nonlinear resistor. This relationship describes the complete device operation over
Figure 1.2 - Examples of constitutive current-voltage relationships for nonlinear resistors:  
  a.) tunnel diode (voltage-controlled),  
  b.) neon bulb (current-controlled),  
  c.) linear resistor (voltage-controlled and current-controlled), and  
  d.) pathological resistor (neither voltage-controlled nor current-controlled).
any swing of voltage or current. The incremental resistance of a nonlinear current-controlled resistor is defined by

\[ R(i) = \frac{dV(i)}{di} \]  \hspace{1cm} (1.3)

where \( v = V(i) \) is the constitutive relationship. Similarly, the incremental conductance of a nonlinear voltage-controlled resistor is defined by

\[ G(v) = \frac{dI(v)}{dv} \]  \hspace{1cm} (1.4)

where \( i = I(v) \) is the constitutive relationship. These quantities are only valid for small signals around a particular operating point.

1.2.2 The nonlinear two-terminal capacitor

A simple (two-terminal) nonlinear capacitor (Fig. 1.3) is similarly described by a constitutive relationship, this time between charge and voltage. The stored charge, \( q \), is the state variable of the capacitor and is related to the branch current by

\[ q = q_o + \int_{-\infty}^{t} i(\tau)d\tau \]  \hspace{1cm} (1.5a)

and

\[ i = \frac{dq}{dt} \]  \hspace{1cm} (1.5b)
Since the device current is the time derivative of the device charge, two capacitors having constitutive charge-voltage relationships which differ only by a constant charge are indistinguishable by terminal current-voltage measurements. In a similar way to the nonlinear resistor, the charge-voltage relationship may be charge-controlled, voltage-controlled or both. The MOSFET is treated, here and by other authors, as a voltage-controlled device; that is, the charges are described by single-valued functions of the terminal voltages. (This should not be confused with 'charge control' analysis, where the currents are described in terms of the device charges.) However, the MOSFET charges are monotonic in the terminal voltages and therefore the device can equally be considered as charge-controlled.
For voltage-controlled devices, we can define an incremental capacitance as follows:

\[ i = \frac{dq}{dt} = \frac{\partial Q(v)}{\partial v} \frac{dv}{dt} = C(v) \frac{dv}{dt} \]  \hspace{1cm} (1.6)

where

\[ C(v) = \frac{\partial Q(v)}{\partial v} \]  \hspace{1cm} (1.7)

and \( q = Q(v) \) is the constitutive relationship. As in the case of the nonlinear resistor, the constitutive charge-voltage relationship is the large-signal description for the nonlinear capacitor, while the capacitance versus voltage relationship is strictly an incremental description which can only be applied at a particular operating point. Therefore, for large-signal transient analysis, a capacitance versus voltage description is less accurate over a voltage transient (given finite numerical precision) than a complete charge versus voltage description. (The implications of the fundamental nature of charge for numerical simulation is discussed further in Section 2.2).

1.2.3 Multi-terminal nonlinear capacitors

The first step in generalizing the nonlinear two-terminal capacitor to \( N \) terminals is to generalize the definite branch notation used above to indefinite nodal variables. (See Fig. 1.4.) The new nodal variables are related to the old branch variables as follows:
\[ v = v_1 - v_2 \]

\[ i = i_1 = -i_2 \quad (1.8) \]

\[ q = q_1 = -q_2 \]

The two terminal charges are related to the two terminal currents by

\[ q_1 = q_{10} + \int_{-\infty}^{t} i_1(\tau) d\tau \quad (1.9a) \]

\[ q_2 = q_{20} + \int_{-\infty}^{t} i_2(\tau) d\tau \]

and

\[ i_1 = \frac{dq_1}{dt} \quad (1.9b) \]

\[ i_2 = \frac{dq_2}{dt} \]

Finally, the voltage-controlled constitutive relationship is transformed as follows:

\[ q_1 = Q(v_1 - v_2) = Q_1(v_1, v_2) \]

\[ q = Q(v) \implies q_2 = -Q(v_1 - v_2) = Q_2(v_1, v_2) \quad (1.10) \]

This notation has two advantages. First, it is consistent with the concept of charge neutrality and charge conservation. For a two-terminal device, charge neutrality
Figure 1.4 - Generalized notation for the two-terminal nonlinear capacitor.

requires that the two charges be equal and opposite. In many cases, these nodal charges will correspond to physically distinct charge populations. For example, a parallel plate capacitor has a charge +q on one plate, which is supplied from the positive terminal, and a charge -q on the other plate, which is supplied from the negative terminal. The same charge neutrality holds for two-terminal semiconductor devices, although the physical location of the charges and their association with a given terminal may not be as clear. The other advantage of this notation is that it is an indefinite nodal description, i.e., there is no commitment to a particular choice of the reference node. This is an advantage in nodal network analysis.
The incremental nodal capacitances for a multi-terminal voltage-controlled capacitor are defined in a similar way to the branch capacitance above. The terminal currents \( i_1 \) and \( i_2 \) can be rewritten as follows:

\[
\begin{align*}
    i_1 &= \frac{dq_1}{dt} = \frac{\partial Q_1(v_1, v_2)}{\partial v_1} \frac{dv_1}{dt} + \frac{\partial Q_1(v_1, v_2)}{\partial v_2} \frac{dv_2}{dt} \\
    &= C_{11} \frac{dv_1}{dt} + C_{12} \frac{dv_2}{dt}, \quad \text{and} \\
    i_2 &= \frac{dq_2}{dt} = \frac{\partial Q_2(v_1, v_2)}{\partial v_1} \frac{dv_1}{dt} + \frac{\partial Q_2(v_1, v_2)}{\partial v_2} \frac{dv_1}{dt} \\
    &= C_{21} \frac{dv_1}{dt} + C_{22} \frac{dv_2}{dt} \quad (1.11)
\end{align*}
\]

where

\[
C_{ij} = \frac{\partial Q_j}{\partial v_i} \quad (1.12)
\]

and the \( C_{ij} \)'s are the terms of the incremental capacitance matrix. From the definitions above, it is clear that \( C_{11} = C_{22} = C \) and \( C_{12} = C_{21} = -C \) for any charge-conserving, reference-independent, two-terminal capacitor description. These incremental capacitances can be interpreted as the change in charge at node \( i \) in response to a change in voltage at node \( j \).

For small-signal sinusoidal excitations this leads to the Indefinite Admittance Matrix (IAM), \( Y(s) \), which is defined by
\[ I(s) = \begin{bmatrix} i_1(s) \\ i_2(s) \end{bmatrix} = Y(s) \begin{bmatrix} v_1(s) \\ v_2(s) \end{bmatrix} = Y(s)V(s) \]  \hspace{1cm} (1.13)

where

\[ Y(s) = s \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \]  \hspace{1cm} (1.14)

is the IAM for the two-terminal capacitor. The IAM is important because it contains a complete, reference-independent, incremental description of the device. For AC circuit simulations, a complete network IAM can be easily constructed from the IAM's of the individual devices in the network. This yields a total network description, relating the network current and voltage vectors, which can be solved for the network node voltages and branch currents.

Extending this notation to multi-terminal devices is quite straightforward. For an N-terminal, voltage-controlled capacitor, the large-signal behavior is described by N charge-voltage relationships of the form

\[ q_i = Q_i(v_1, v_2, v_3, \ldots, v_N) \]  \hspace{1cm} (1.15)

Similarly, the small-signal behavior can be described by the \( N^2 \) incremental capacitances, \( C_{ij} \), which form the IAM. These terms can be calculated from (1.12) as above. For physical devices, the charge-voltage relationships must be defined self-consistently such that the rows and columns of the IAM sum to zero.
\[ \sum_{i=1}^{N} C_{ij} = 0 \quad \text{and} \quad \sum_{j=1}^{N} C_{ij} = 0 \quad \text{(1.16)} \]

Each row must sum to zero for the IAM to be reference-independent, and each column must sum to zero for the device description to be charge conservative, which is equivalent to obeying KCL. As a result, of the \( N^2 \) capacitance terms only \( N^2-(2N-1) \) are independent. (One of the \( 2N \) constraints is redundant.) For example, the MOSFET, which is a four-terminal device, has 16 incremental capacitance terms of which only 9 are independent.

An alternative notation for the incremental capacitances is frequently used when referring to the capacitances of a specific device, rather than the general capacitance matrix terms. In this case, the small signal device capacitances are defined by

\[ C_{ij} = \frac{\partial Q_i}{\partial v_j} \quad \text{if} \quad i=j \]
\[ C_{ij} = -\frac{\partial Q_i}{\partial v_j} \quad \text{if} \quad i\neq j \quad \text{(1.17)} \]

where the signs of the \( C_{ij} \)'s are chosen to keep all or the capacitance terms positive for well-behaved devices, i.e., devices for which the charge at a node increases with an increase in the voltage at that node and decreases with an increase in the the voltage at any other node. The IAM for

- 34 -
a two-terminal capacitor using this convention would be written as

\[ \mathbf{Y}(s) = s \begin{bmatrix} C_{11} & -C_{12} \\ -C_{21} & C_{22} \end{bmatrix} \quad (1.18) \]

where \( C_{11} = C_{12} = C_{21} = C_{22} = C \) for a charge-conserving, reference-independent, two-terminal capacitor description. This notation will be used in Chapter 2 when the MOSFET small signal capacitances are discussed.

1.2.4 Reciprocity and conservation of energy

So far nothing has been said about the reciprocity or nonreciprocity of the multi-terminal capacitor. In this context, a multi-terminal capacitor model is said to be reciprocal at an operating point if the IAM evaluated at that operating point is symmetric, that is, if \( C_{ij} = C_{ji} \) for all \( i \) and \( j \). The model is said to be reciprocal if it is reciprocal at every operating point.

The simplest example of a reciprocal capacitor is the two-terminal capacitor discussed above. It is easily shown that any two-terminal capacitor which is charge-conserving and reference-independent is reciprocal. For the two-terminal case, only one of the four incremental capacitances is independent and, therefore, all of the incremental capacitances must be equal and the IAM is symmetric. It is also
an elementary property of simple, two-terminal, linear capacitors that they are lossless. This means that the net power generated or dissipated as the terminal voltages move through any closed path is zero. In the general case, a nonlinear multi-terminal capacitor is lossless if and only if the IAM is symmetric at every operating point. (A recent proof of this has been provided by Wyatt et al in [1.4,1.5].) So, the property of capacitors to be lossless does not depend on the number of terminals or on linearity but only on reciprocity.

As an example of a nonreciprocal multi-terminal capacitor, consider the transcapacitor shown in Fig. 1.5, which is defined by the following constitutive charge-voltage relationship:

\[ q_1 = C (v_1 - v_3) \]
\[ q_3 = C (v_3 - v_1) \]  \hspace{1cm} (1.19)
\[ q_2 = 0 \]

This element is analogous to the DC transconductor in that the control voltage is not the same as the branch voltage. The IAM for the transcapacitor is constructed by calculating the nine incremental capacitance terms, again using (1.12). The result is shown in Fig. 1.6. Notice that this IAM is nonreciprocal. For example, if the voltage at node 3 is excited, the current observed leaving node 1 is \( C \frac{dv_3}{dt} \).
Figure 1.5 - The linear transcapacitor.

\[
\begin{bmatrix}
C & 0 & -C \\
-C & 0 & C \\
0 & 0 & 0
\end{bmatrix}
\]

\(Y(s) = s\)

Figure 1.6 - The IAM for the linear transcapacitor.
However, if the voltage at node 1 is excited, the current observed leaving node 3 is zero. So, $C_{11}$ is zero while $C_{12}$ is non-zero. This nonreciprocity arises from the separation of the control voltage and the branch voltage as noted above. It is demonstrated below that this separation also leads to an unbounded generation of power with the proper choice of control and branch voltages.

Consider the voltage waveforms for $v_{11}$ and $v_{12}$ shown in Fig. 1.7. Both the control and branch voltages move through a closed path in a time interval $T$. The instantaneous power dissipation, $P_d$, is given by the product of the branch voltage ($v_{11}$) and the branch current ($C \frac{dv_{11}}{dt}$) and is also shown in Fig. 1.7. The total power dissipated over the interval $T$ is non-zero and negative. In other words, as the terminal voltages move through the closed path, the transcapacitor generates net energy equal to $C$ joules (the area under the $P_d$ curve.) Clearly, if this experiment is repeated at a faster and faster rate (decreasing $T$) it is possible to generate an unbounded amount of energy per unit time.

This example demonstrates that a charge-conserving, reference-independent, multi-terminal capacitor model can be nonreciprocal, and if such a model is nonreciprocal, then it can be driven in such a way as to generate an unbounded amount of power. Therefore, a capacitor model for the
Figure 1.7 - Demonstration of non-conservation of energy for the linear transcapacitor.
MOSFET which is nonreciprocal, such as those proposed by Ward [1.1,1.2] and others [1.6-1.10], is not completely satisfying. However, it is clear from the demonstration above that the power generated by a nonreciprocal capacitor is unbounded only if the frequency of excitation is unbounded. So, a nonreciprocal capacitor model may well be appropriate for physical devices over a restricted frequency range.

1.2.5 Experimental capacitance measurements of multi-terminal devices

As discussed above, a multi-terminal capacitor can be fully described only by its constitutive charge-voltage relationship. Therefore, for a large-signal characterization of a multi-terminal capacitor, the charge-voltage relationship must be reconstructed from current-voltage measurements. Conceptually, this can be done by starting with the device at some initial bias condition and then integrating the terminal currents as the terminal voltages are moved through all of the bias conditions of interest. This will produce the complete charge-voltage relationship to within a constant associated with the charges at the initial bias condition. In practice, this approach is extremely difficult since any errors in the terminal current measurements (such as an offset current) will be integrated throughout the exercise. This approach is sometimes used
for (two-terminal) MOS capacitors, but generally only the calculated incremental capacitances, which do not reflect the integrated errors, are used.

More commonly, multi-terminal capacitors are experimentally characterized by measuring the individual capacitance terms of the incremental IAM, and plotting these capacitances against the terminal voltages. This information is frequently used to confirm a hypothetical large-signal charge-voltage relationship. The individual capacitance terms are obtained by taking the imaginary (or out-of-phase) part of two-terminal, small-signal sinusoidal admittance measurements. These two-terminal admittances are measured by applying a small AC voltage at one terminal of the device and measuring the amplitude and phase of the AC current at another terminal. (A capacitance instrument is simply an admittance instrument which ignores the real part of these admittances.) Most instruments cannot measure the diagonal elements of the IAM, the $C_{ii}$'s, which represent a one-point measurement. However, since the $N^2$ terms of the IAM are not all independent, measuring the off-diagonal terms is, in principle, sufficient.

It is important when characterizing real devices to maintain a distinction between the physical device and the model. The admittance measurements described above will yield valid incremental capacitances only if the device
being tested can reasonably be modeled as a multi-terminal capacitor in parallel with a multi-terminal resistor. For example, consider a linear transmission line. The admittances of a linear transmission line are infinite series in $s$ (complex frequency). The IAM can be generalized to a matrix of infinite series, but the multi-terminal resistor and capacitor defined above cannot. In particular, at any single frequency the real and imaginary parts of the series IAM can be calculated. At that one frequency, therefore, the transmission line can be modeled as a multi-terminal resistor in parallel with a multi-terminal capacitor. However, if the signal frequency changes, a new resistor and capacitor model is required. It would be tempting to construct a frequency-dependent resistor and capacitor model from the frequency-dependent real and imaginary parts of the IAM terms; however, there is no large-signal constitutive relationship which would match such an incremental model.

As another example, the Ward nonreciprocal capacitance model has been shown to agree with the experimental low-frequency small-signal capacitances of the MOSFET [1.11,1.12]. However, it is nonsensical to assert that the MOSFET is a nonreciprocal capacitor, since this would imply the capability for infinite power generation. At best, it can be said that the MOSFET is well modeled by a nonreciprocal capacitor at low frequencies. The point to keep in mind is that, while admittance measurements can always be
made, it is not always appropriate to interpret the imaginary part of these measurements as the incremental capacitances of a multi-terminal capacitor model.
CHAPTER 2

History of MOSFET Dynamic Modeling

This chapter reviews the history of both quasi-static and non-quasi-static dynamic modeling of the MOSFET. The most historically significant quasi-static model is the model formulated by Meyer in 1971 [2.1] which has been extensively used in computer simulators such as SPICE [2.2]. This model has recently been replaced by the Ward model [2.3,2.4], which addresses several fundamental problems with the Meyer formulation. The relationship between these two models will be discussed in depth, with particular emphasis on the phenomenon of nonreciprocal capacitance in the Ward model.

Several non-quasi-static models of the MOSFET were also developed in the late 1960's; however, these models have rarely been used for practical circuit analysis. A non-quasi-static analysis is presented here which gives some further insight into the relationship between the Meyer and Ward quasi-static capacitance models. This analysis produces an admittance matrix for the device in terms of infinite series in $s$ (complex frequency). The Meyer and Ward capacitances can be identified as different approximations to the admittance series.

- 45 -
2.1 Static Analysis of the MOSFET

This section reviews the static modeling of the MOSFET. First, the application of the fundamental semiconductor equations to the MOSFET is described, and then, as a specific example of this analysis, the static transport current is calculated for the case of strong inversion. This material is based on the method of analysis first presented by Inntonta and Moll [2.5], and subsequently refined by several authors in standard texts [2.6-2.8]. The development and notation used here closely follows that of Ward in [2.9]. It is assumed that the reader is familiar with the basic concept of metal-oxide-semiconductor systems, i.e., the accumulation, depletion or inversion of a semiconductor surface by the application of a vertical field.

2.1.1 General analysis of the MOSFET

The physical structure of a n-channel MOSFET is shown in Fig. 2.1. The origin of the coordinate system is chosen as the silicon surface at the source end of the channel region. The x-axis extends from the surface (x=0) into the substrate, while the y-axis lies along the channel from the source (y=0) to the drain (y=L). The device is assumed to be uniform in the z-axis with a width of W.
The MOSFET, like any semiconductor device, can be described by three basic equations: the carrier transport and continuity equations and Poisson's equation. In the case of the n-channel MOSFET, the device current is assumed to consist only of electrons flowing in a thin inversion layer near the surface. Hole currents are neglected, and it is assumed that there are no currents perpendicular to the silicon surface, i.e., generation currents in the depletion region are ignored.

With these assumptions, the electron current density, which is $y$-directed (along the channel), can be described by

$$J_y = -q \mu_n n \frac{\partial \psi}{\partial y} + qD_n \frac{\partial n}{\partial y}$$  \hspace{1cm} (2.1)
where \( n \) is the mobile electron concentration, \( \mu_n \) is the electron mobility, \( D_n \) is the electron diffusivity and \( \psi \) is the electrostatic potential. Given Boltzmann statistics, and making use of the Einstein relation \( \mu_n = (q/kT) D_n \), the electron current density can be rewritten as

\[
J_y = - q \mu_n n \frac{\partial \phi_n}{\partial y}
\]  

(2.2)

where \( \phi_n \) is the quasi-Fermi potential for electrons, defined by

\[
n = n_i e^{(\psi - \phi_n)/kT}
\]  

(2.3)

The total channel current, as a function of \( y \), is found by integrating (2.2) in the \( x \) and \( z \) directions. This produces

\[
I(y,t) = - W \int_0^\infty q \mu_n n \frac{\partial \phi_n}{\partial y} dx = W \mu(y,t) Q_n(y,t) \frac{\partial V(y,t)}{\partial y}
\]  

(2.4)

where

\[
Q_n(y,t) = - q \int_0^\infty n(x,y,t) dx
\]  

(2.5)

is the channel charge density per unit area, and

\[
V(y,t) = \phi_n(x,y,t) \bigg|_{x=0}
\]  

(2.6)

is the quasi-Fermi potential at the surface. The reference for \( V(y,t) \) is chosen such that \( V(0) = V_S \) and \( V(L) = V_D \).
Strictly speaking, the upper limit on the integral in (2.5) should be chosen to include all of the electrons in the inversion layer but not the electrons deep in the bulk. This construction is not conceptually difficult since these two electron populations are separated by a large depletion region. Also, it has been assumed that there are no vertical currents, which, from (2.2), implies that $\partial \phi_n / \partial x$ must be zero at least near the surface where $n$ is appreciable. Therefore, the $\partial \phi_n / \partial y$ term is independent of $x$ and can be pulled out of the integral. This assumption also justifies the replacement of $\partial \phi_n$ by $\partial V(y,t)$. Finally, $\mu_n$ has also been pulled out of the integral and replaced by the effective channel mobility $\mu(y,t)$. Although the relationship between $\mu_n$ and $\mu(y,t)$ is not made explicit there is no loss of generality. Using similar notation, the current continuity equation can be written as

$$\frac{\partial I(y,t)}{\partial y} = -W \frac{dQ_n(y,t)}{dt}.$$  \hspace{1cm} (2.7)

For large-geometry devices, the application of Poisson’s equation can be simplified by the Gradual Channel Approximation (GCA). The GCA asserts that the perpendicular field at the silicon surface is much larger than the lateral field, and therefore Poisson’s equation can be reduced to one dimension. For this case, the mobile channel charge at a point $y$ along the channel can be written in terms of the
surface potential (or the quasi-Fermi potential) at that point, along with the gate and bulk voltages. In other words we can write

\[ Q_n(y,t) = Q_n(V(y,t), V_G(t), V_B(t)) = Q_n(V, V_G, V_B) \quad (2.8) \]

A similar assumption applied to the channel mobility produces

\[ \mu(y,t) = \mu(V(y,t), V_G(t), V_B(t)) = \mu(V, V_G, V_B) \quad (2.9) \]

(Note that (2.9) will not apply in the high-field drain region of a long channel device in saturation, where the mobility will also depend upon the lateral electric field.) Incorporating (2.8) and (2.9) into (2.4) produces the revised form of the transport equation;

\[ I(y,t) = W \mu(V, V_G, V_B) Q_n(V, V_G, V_B) \frac{dV}{dy} \quad (2.10) \]

From the current continuity equation, (2.7), it is clear that the static transport current, \( I_0 \), is constant along the channel, i.e., \( \frac{dI(y,t)}{dy} = 0 \) when \( \frac{dQ_n}{dt} = 0 \). Therefore, \( I_0 \) can be found by integrating (2.10) from source to drain as shown below:

\[ I_0 = \frac{W}{L} \int_{V_S}^{V_D} \mu(V, V_G, V_B) Q_n(V, V_G, V_B) dV \quad (2.11) \]
To solve (2.11), a solution must be found to the one-dimensional Poisson's equation which will relate the mobile channel charge density, $Q_n$, and the surface quasi-Fermi potential, $V(y)$.

### 2.1.2 The strong inversion approximation

As an example of a particular solution for the static transport current, consider the case of constant mobility and strong inversion. From (2.11), constant mobility implies

$$I_o = \mu_L \int_{V_S}^{V_D} Q_n(V, V_G, V_B) dV$$  \hspace{1cm} (2.12)

For strong inversion it is assumed that the surface potential, $\psi_S(y)$, is fixed at $V(y) + \phi_F$, where

$$\phi_F = \frac{kT}{q} \ln \frac{N_A}{n_i}$$  \hspace{1cm} (2.13)

is the quasi-Fermi potential deep in the bulk. This means that the surface potential is shifted relative to its value deep in the bulk, $\psi_B$, by

$$\psi_S(y) - \psi_B = V(y) - V_B + 2\phi_F$$  \hspace{1cm} (2.14)

where $\psi_B = V_B - \phi_F$ relates the electrostatic and quasi-Fermi potentials in the bulk. (As in the case of $V(y)$, the refer-
ence for the quasi-Fermi potential in the bulk is chosen to be the terminal potential, \( V_B \). This shift of electrostatic potential is referred to as band bending, and it is equal to the potential across the channel-bulk depletion region. This can be seen in the band diagram of Fig. 2.2, which shows the energy levels of the conduction band \( (E_C) \), the valence band \( (E_V) \), the intrinsic level \( (E_i) \), plus the quasi-Fermi levels for electrons at the surface \( (E_{fn}) \) and holes in the bulk \( (E_{fp}) \).

\[ q(\psi_S - \psi_B) \]

\[ E_{fn} \]

\[ q(V - V_B) \]

\[ q\phi_F \]

\[ E_{fp} \]

\[ E_i \]

\[ E_C \]

\[ E_V \]

**Figure 2.2** - Band diagram of the MOSFET in strong inversion.
At this point it is easy to calculate the gate and bulk charge densities, \( Q_g(y) \) and \( Q_b(y) \). The gate charge is given by \( C_{ox}V_{ox} \), where \( C_{ox} \) is the gate capacitance per unit area and \( V_{ox} \) is the voltage across the oxide (from the gate to the silicon surface.) In the absence of a work function difference, \( V_{ox} \) is given by the gate-to-bulk bias minus the band bending at the surface. Therefore,

\[
Q_g(y) = C_{ox}[(V_G - V_B) - (V(y) - V_B + 2\phi_F)] .
\]  

(2.15)

Work function differences and fixed oxide charges can be accounted for by replacing \( V_G \) with \( V_g = V_G - V_{FB} \), where \( V_{FB} \) is the flat-band voltage. Using this notation, (2.15) can be simplified to

\[
Q_g(y) = C_{ox}(V_g - V(y) - 2\phi_F) .
\]  

(2.16)

The bulk charge density can be calculated using the depletion approximation as follows:

\[
Q_b(y) = -C_{ox}\gamma[V(y) - V_B + 2\phi_F]^{1/3}
\]

(2.17)

where

\[
\gamma = \frac{(2\varepsilon_S qN_A)^{1/3}}{C_{ox}} .
\]

(2.18)

Finally, since a one-dimensional Poisson's equation is used in vertical slices along the channel, charge neutrality must also hold at each point along the channel. Therefore, the mobile channel charge density, \( Q_n(y) \), can be found from
\[ Q_n(y) + Q_g(y) + Q_b(y) = 0 \quad (2.19) \]

so that

\[ Q_n(V) = -C_{ox} \left[ (V_g - V - 2\phi_P) - \gamma (V - V_B + 2\phi_P)^{\frac{3}{2}} \right] . \quad (2.20) \]

Incorporating (2.20) into (2.12) and integrating from source to drain produces

\[ I_o = \frac{\mu C_{ox} W}{L} \left[ (V_g - V_B) V_{DS} - \frac{2}{3} \gamma \left[ (V_{DB} + 2\phi_P)^{\frac{3}{2}} - (V_{SB} + 2\phi_P)^{\frac{3}{2}} \right] \right. \]
\[ \left. - \frac{1}{2} \left[ (V_{DB} + 2\phi_P)^2 - (V_{SB} + 2\phi_P)^2 \right] \right] . \quad (2.21) \]

This analysis for the static transport current only applies as long as there is a continuous channel from source to drain. If the drain voltage is increased beyond the saturation voltage, \( V_{DSat} \), the channel will be pinched off at the drain end. The drain voltage in excess of \( V_{DSat} \) will be dropped across a high field drain region, and the potential at the pinched-off end of the channel will be fixed at \( V_{DSat} \). The analysis for the channel remains unchanged except for the use of this pseudo drain voltage. Therefore, the expression for \( I_o \) above remains valid in saturation if \( V_D \) is replaced by \( V_{DSat} \), where \( V_{DSat} \) is calculated from

\[ Q_n(L) = -C_{ox} \left[ (V_g - V_{DSat} - 2\phi_P) - \gamma (V_{DSat} - V_B + 2\phi_P)^{\frac{3}{2}} \right] = 0 \quad (2.22) \]

Modulation of the width of the drain region by the excess drain voltage produces a non-zero output conductance; the
transport current is not exactly constant in saturation for increasing $V_{DS}$. However, this effect is generally modeled by a bias dependence in the channel length, $L$, rather than by changes in the analysis above.

An important feature of this analysis is that an explicit calculation of $V(y)$ is not required. This will also be true of the quasi-static capacitance calculations in the next section. The shape of $V(y)$ as it varies along the channel from $V_S$ to $V_D$ changes as the bias conditions change. It is constrained by (2.4); as $V(y)$ increases from source to drain, the mobile charge density decreases and therefore, the slope of $V(y)$ must increase in order to keep the static transport current constant. For short-channel devices the lateral field can become so large that the velocity of the carriers saturates before $V_{Dsat}$ is reached. This is called velocity saturation, and it will be an important consideration in Chapter 4.
2.2 The Meyer Quasi-Static Capacitance Model

This section reviews the Meyer and other early quasi-static capacitance models. These models are based upon the total gate and bulk charges, which can be calculated from the gate and bulk charge densities derived above. Since the gate and bulk charges, and the total channel charge, are insufficient to formulate a complete nonlinear capacitance model as described in Section 1.2, the Meyer model and its predecessors are inherently small-signal models. These models attempt to represent the MOSFET small-signal capacitances without complete information about the terminal charges, in particular, without any knowledge of the individual source and drain charges. As a result, conservation of charge problems arise when the Meyer model is used for circuit simulation. This effect is also discussed below.

2.2.1 Calculation of the total gate and bulk charges

The static analysis in Section 2.1 makes use of the gate and bulk charge densities along the channel, \( Q_g(y,t) \) and \( Q_b(y,t) \), expressed in terms of the surface quasi-Fermi potential, \( V(y,t) \). The total gate and bulk charges are calculated by integrating from source to drain as follows:

\[
Q_G = W \int_0^L Q_g(V, V_G, V_B) \, dy
\]

(2.23)
and

\[ Q_B = W \int_{0}^{L} Q_b(V, V_G, V_B) \, dy \]  \hspace{1cm} (2.24)

These integrals can be carried out by noting from (2.10) that

\[ dy = - \frac{W}{I(y,t)} \mu(V, V_G, V_B) Q_n(V, V_G, V_B) \, dV \]  \hspace{1cm} (2.25)

For a quasi-static analysis we can replace \( I(y,t) \) with \( I_0 \) in (2.25) to get

\[ Q_G = - \frac{W^2}{I_0} \int_{V_S}^{V_D} Q_g(V, V_G, V_B) \mu(V, V_G, V_B) Q_n(V, V_G, V_B) \, dV \]  \hspace{1cm} (2.26)

and

\[ Q_B = \frac{W^2}{I_0} \int_{V_S}^{V_D} Q_b(V, V_G, V_B) \mu(V, V_G, V_B) Q_n(V, V_G, V_B) \, dV \]  \hspace{1cm} (2.27)

The total channel charge can be calculated from a similar integral, or by charge neutrality from

\[ Q_N = -(Q_G + Q_B) \]  \hspace{1cm} (2.28)

These integrals can be evaluated using any particular solution for \( Q_g(V) \), \( Q_b(V) \), and \( Q_n(V) \). The resulting charges are quasi-static because of the substitution of \( I_0 \) for \( I(y,t) \) in (2.25). By quasi-static we mean that the charges are functions only of the instantaneous terminal voltages.
and not their time derivatives. In other words, the internal distributions of potential and charge are assumed to be always identical to their static solutions for the instantaneous terminal voltages. The difference between the Meyer [2.1] and other early quasi-static models is the way these charges are used to formulate a capacitance model.

2.2.2 The Meyer quasi-static capacitance model

The earliest quasi-static capacitance model based only on the gate and bulk charges was presented by Ihantola and Moll in [2.5]. This model was a two-port incremental model of the MOSFET used in the common source configuration. This model included an input capacitance (at the gate), an output capacitance (at the drain) and a feedback capacitance from drain to gate. The output and feedback capacitances were assumed to be equal to the parasitic junction and overlap capacitances respectively, while the input capacitance was calculated by $C_{in} = \frac{\partial Q_N}{\partial V_G}$. (Actually, $\frac{\partial Q_G}{\partial V_G}$ would have been a better choice for the input capacitance.) Although the bulk charge was used to calculate the input capacitance, no capacitances to the substrate were included.

Several similar models were presented in the late 1960's [2.10,2.11], many of which were based on earlier JFET equivalent circuit models [2.12-2.14]. These models generally share the same derivation; a two-port circuit-
equivalent model is assumed, and then an attempt is made to calculate the values of the capacitance elements from the gate and bulk charges. Frequently, the equivalent circuits were taken from non-quasi-static models, so the relationship between the capacitance elements of the model and the physical device charges is tenuous at best.

The first complete capacitance model based only on gate and bulk charges was presented by Cobbold [2.8]. A simplified formulation of this model was later proposed by Meyer for use in circuit simulators [2.1]. Meyer's work was apparently independent, and, although Cobbold's work appeared first and was more complete, this general approach to MOSFET capacitance modeling has historically been attributed to Meyer. This has occurred because the Meyer paper was referenced by the authors of SICE [2.2], which later became the most widely used circuit simulator for MOS circuits. Accordingly, this approach is referred to here as the Meyer model and no distinction is made between the more complete discussion of [2.8] and the simpler form of [2.1]. For example, although substrate capacitances were included by Cobbold and ignored by Meyer they are included in the discussion below under the label of the Meyer model. In any case, the important contribution of the Meyer model was that it provided an indefinite, multi-terminal description of the MOSFET capacitances (rather than a specific two-port description) which was related directly to the gate and bulk

- 59 -
The basic concept of the Meyer model is an assertion about the source and drain terms of the Indefinite Admittance Matrix (IAM).

Consider the IAM of the Meyer capacitance model shown in Fig. 2.3. The gate capacitances (top row) and the bulk capacitances (bottom row) can be calculated directly from the gate and bulk charges using (1.17). However, without an individual calculation of source and drain charges, the source and drain capacitances can not be similarly calculated to complete the matrix. The Meyer formulation simply asserts that the IAM must be symmetric and therefore the source and drain terms can be filled in from the gate

\[
\begin{bmatrix}
C_{gg} & -C_{gs} & -C_{gd} & -C_{gb} \\
-C_{gs} & (C_{gs} + C_{bs}) & 0 & -C_{bs} \\
-C_{gd} & 0 & (C_{gd} + C_{td}) & -C_{bd} \\
-C_{bg} & -C_{bs} & -C_{bd} & C_{bb}
\end{bmatrix}
\]

Figure 2.3 - The Meyer capacitance matrix.

- 60 -
and bulk capacitances. In addition, it is assumed that there is no capacitance between the source and drain.

This procedure only works if the gate-to-bulk and bulk-to-gate capacitances \( (C_{gb} \text{ and } C_{bg}) \) calculated from (1.17) are equal. If they are not, the gate and bulk columns of Fig. 2.3 will not sum to zero and the matrix will not obey KCL. Indeed, these capacitances are generally not equal. The IAM can be resolved by using \( C_{gb} \) in place of \( C_{bg} \), and replacing \( C_{bb} \) by \( (C_{gb} + C_{bs} + C_{bd}) \), as shown in Fig. 2.4. Since a symmetric IAM, like this one, can be represented by a network of simple, two-terminal capacitors, the Meyer IAM is equivalent to the all-capacitor network of Fig 2.5.

\[
\begin{bmatrix}
C_{gg} & -C_{gs} & -C_{gd} & -C_{gb} \\
-C_{gs} & (C_{gs} + C_{bs}) & 0 & -C_{bs} \\
-C_{gd} & 0 & (C_{gd} + C_{bd}) & -C_{bd} \\
-C_{gb} & -C_{bs} & -C_{bd} & (C_{gb} + C_{bs} + C_{bd})
\end{bmatrix}
\]

Figure 2.4 - The resolved Meyer capacitance matrix.
Figure 2.5 - Equivalent circuit for the Meyer capacitance model.
Actually, Cobbold and Meyer developed this approach from an equivalent circuit point of view. The assertion of reciprocity of the IAM is equivalent to assuming an all-capacitor network. However, the discussion above makes it clear that the Meyer model makes the best possible choice of these capacitor values. Cobbold recognized the nonreciprocity of the $C_{gb}$ and $C_{bg}$ terms of the IAM and attempted to account for it in his equivalent circuit model by adding a transcapacitor. However, his equivalent circuit model did not correctly implement the IAM. It should be clear from the IAM derivation above that if the source and drain capacitance terms are reciprocal, the gate-bulk capacitance pair must also be equal. So, Cobbold's attempt to retain the nonreciprocity of the IAM was not consistent with charge conservation (or KCL) and reference-independence.

2.2.3 Non-conservation of charge in circuit simulators

As mentioned above, the Meyer formulation provides an incremental capacitance model and not a charge-based large-signal model. As a result, the discretization in time of the Meyer model for transient analysis can only be based on capacitance and not charge. This causes numerical integration errors which can lead to non-conservation of charge. This effect was discussed in a text by Chua and Lin [2.15] on numerical circuit analysis techniques in 1975. However, the practical problems experienced with using the Meyer
model in simulators like SPICE were not really recognized until Ward did so in 1978 [2.3]. A more thorough analysis of this problem has recently been made by Yang et al [2.16]. A qualitative discussion of this problem is given below.

Consider the branch current of a two-terminal nonlinear capacitor as described by its small-signal capacitance

$$i = C(v) \frac{dv}{dt} \quad . \quad (2.29)$$

Now consider the integration of (2.29) over a time step interval from \( t_n \) to \( t_{n+1} \):

$$\int_{t_n}^{t_{n+1}} i dt = \int_{v_n}^{v_{n+1}} C(v) dv \quad . \quad (2.30)$$

The left side of (2.30) can be discretized and evaluated by

$$\int_{t_n}^{t_{n+1}} i dt \approx \frac{h}{2} (i_{n+1} + i_n) \quad (2.31)$$

where the trapezoidal rule for numerical integration has been used and \( h = (t_{n+1} - t_n) \) is the time step. The error in (2.31) comes from the discretization of \( i(t) \) in time and is fundamental to numerical analysis. The error of interest here comes from the evaluation of the right side of (2.30), which is evaluated as follows:
\[ \tau_{n+1} \int_{t_n}^{t_{n+1}} \frac{v_{n+1}}{C(v)dv} = \tilde{C}(v_{n+1} - v_n) \]  

(2.32)

where \( \tilde{C} \) is some average capacitance over the interval from \( t_n \) to \( t_{n+1} \). The average capacitance is generally evaluated at either \( v_n \) (explicit) or \( v_{n+1} \) (implicit), but in either case there will be some error. It can be shown that this error can accumulate over a voltage transient, producing an apparent loss (or gain) of charge [2.3, 2.15, 2.16].

In contrast, consider the branch current of a two-terminal capacitor as described by the large-signal charge relationship

\[ i = \frac{dq(v)}{dt} . \]  

(2.33)

When (2.33) is integrated over a time step, the right side of the equation is evaluated by

\[ \int_{t_n}^{t_{n+1}} \frac{q_{n+1}}{dq} = q(v_{n+1}) - q(v_n) \]  

(2.34)

which is an exact expression in terms of \( v_n \) and \( v_{n+1} \). This demonstrates the importance of using a charge-based model for large-signal analysis, rather than a small-signal capacitance model.
2.3 The Ward Model and Other Charge Partitioning Models

The stumbling block in deriving a large-signal model for the MOSFET capacitances is the calculation of the individual source and drain charges. Essentially, a partitioning algorithm for the total channel charge is needed to separate the two terminal charges. It is this charge partitioning problem that was first addressed by Ward, with Oh and Dutton, in [2.4]. This analysis is presented below along with several of the charge partitioning schemes of other authors. In addition, a circuit equivalent model, which includes transcapacitors, is presented for the nonreciprocal Ward IAM matrix. This model gives some circuit intuition into the MOSFET capacitance characteristics.

2.3.1 The Ward model

The problem of non-conservation of charge which arises from the use of small-signal MOSFET capacitance models for large-signal analysis was first pointed out by Ward and Dutton in 1978 [2.3]. The charge partitioning scheme proposed in [2.3] was simply to assign half of the channel charge to the source and half to the drain. This simple partitioning was shown to lead to nonreciprocal small-signal capacitances. However, no experimental data were presented in [2.3], nor was there a justification for the arbitrary 50-50 partitioning algorithm.
In a 1980 paper authored by Oh, Ward and Dutton [2.4] a new charge partitioning scheme was presented which was derived from a physical analysis of the MOSFET. A more complete discussion of this analysis was presented by Ward in [2.9], including an experimental capacitance characterization of a large device which supported this model. This model is referred to here as the Ward model, and its derivation is reviewed below.

First, the current continuity equation, (2.7) above, is integrated from the source \((y=0)\) to an arbitrary point \(y\) along the channel producing

\[
I(y,t) - I_S(t) = - W \int_0^y \frac{\partial Q_n(y',t)}{\partial t} \, dy' \quad (2.35)
\]

where \(I_S(t)\) is the current at the source. Now, \(I(y,t)\) can be replaced using the transport equation, (2.4), to produce

\[
I_S(t) = W\mu(y,t)Q_n(y,t) \frac{\partial V(y,t)}{\partial y} + W \int_0^y \frac{\partial Q_n(y',t)}{\partial t} \, dy' \quad (2.36)
\]

Integrating (2.36), this time from source to drain, produces

\[
I_S(t) = \frac{W}{L} \int_0^L \mu(y,t)Q_n(y,t) \frac{\partial V(y,t)}{\partial y} \, dy
\]

\[
+ \frac{W}{L} \int_0^L \int_0^y \frac{\partial Q_n(y',t)}{\partial t} \, dy' dy \quad (2.37)
\]

- 67 -
The last term in (2.37) can be rewritten to give

\[ I_s(t) = \frac{W}{L} \int_0^L \mu(y,t)Q_n(y,t) \frac{\partial V(y,t)}{\partial y} \, dy \]

\[ + \frac{d}{dt} \left[ W \int_0^L (1 - \frac{y}{L})Q_n(y,t) \, dy \right]. \quad (2.38a) \]

Following the same procedure, this time integrating (2.7) from a point \( y \) to the drain \( (y=L) \), produces a similar equation for the drain current

\[ I_d(t) = -\frac{W}{L} \int_0^L \mu(y,t)Q_n(y,t) \frac{\partial V(y,t)}{\partial y} \, dy \]

\[ + \frac{d}{dt} \left[ W \int_0^L \frac{y}{L}Q_n(y,t) \, dy \right]. \quad (2.38b) \]

Therefore, \( I_s(t) \) and \( I_d(t) \) can be written as

\[ I_s(t) = -I_T(t) + \frac{dQ_s(t)}{dt} \quad (2.39a) \]

\[ I_d(t) = I_T(t) + \frac{dQ_d(t)}{dt} \quad (2.39b) \]

where

\[ I_T(t) = -\frac{W}{L} \int_0^L \mu(y,t)Q_n(y,t) \frac{\partial V(y,t)}{\partial y} \, dy \quad (2.40) \]

is the transport current and
\[ Q_s = W \int_0^L (1-\frac{y}{L}) Q_n \, dy \quad (2.41a) \]

and

\[ Q_d = W \int_0^L \frac{y}{L} Q_n \, dy \quad (2.41b) \]

are the source and drain charges. The source and drain charges can be evaluated in a similar way to the gate and bulk charges using (2.25) to replace \( dy \). The \( y/L \) terms in (2.41) can be evaluated by integrating (2.25) from the source to a point \( y \) along the channel. Note that in general the transport current, \( I_T(t) \), is a function of time and need not be quasi-static. However, in the gradual channel case, (2.40) becomes identical to (2.11), so that \( I_T(t) = I_0 \) as calculated in section 2.1.1.

Given the partitioning functions of (2.41), all four terminal charges can be calculated in terms of the terminal voltages. This produces a complete large-signal, nonlinear capacitance model for the MOSFET, and each of the 16 small-signal capacitances can be calculated directly from (1.17). The resulting IAM will in general be nonreciprocal. As an example of the nonreciprocal capacitances that arise from the Ward model, Figs. 2.6-2.8 are reproduced from [2.9]. Fig. 2.6 shows \( C_{gs} \) and \( C_{sg} \) versus gate voltage with drain-to-source voltage as a parameter. The solid lines are
Figure 2.6 - a.) $C_{gs}$, and b.) $C_{sg}$ for a large-geometry device (from [2.3]).
Figure 2.7 – a.) $C_{gd}$, and b.) $C_{dg}$ for a large-geometry device (from [2.3]).
Figure 2.8 - a.) $C_{sd}$, and b.) $C_{ds}$ for a large-geometry device (from [2.3]).
calculated, using the same strong inversion approximation as Section 2.1, and the dotted curves are actual data points. These data are taken from a 100 \mu m by 100 \mu m transistor. The experimental data are in good agreement with the calculated capacitances, except for low gate voltages where the strong inversion approximation is in error. Fig. 2.7 shows \( C_{gd} \) and \( C_{dg} \) for the same bias conditions, and Fig. 2.8 shows \( C_{sd} \) and \( C_{ds} \). The \( C_{sd}, C_{ds} \) pair are interesting because, in addition to being nonreciprocal, these capacitances are negative.

These nonreciprocal and negative capacitances are not really counterintuitive. For example, consider \( C_{gd} \) and \( C_{dg} \) when the device is in saturation. \( C_{gd} \) represents the change in gate charge for a change in drain voltage. In saturation, the drain voltage is isolated from the drain end of the channel, which is fixed at \( V_{Dsat} \), and therefore \( C_{gd} \) is zero. However, as the gate voltage varies, the total channel charge changes. If any of this charge is supplied from the drain, \( C_{dg} \) will be non-zero. As another example, consider \( C_{sd} \) when the device is biased with \( V_{DG} = 0 \). If the drain voltage increases slightly, the total channel charge will increase, i.e., the total number of mobile electrons in the channel will decrease. Since the device is biased symmetrically, some of this increase in charge will be supplied by the source, and if the source supplies positive
charge while the drain voltage increases, a negative capacitance is observed.

One problem with the Ward model, however, is the complexity of the calculated source and drain charges, and of the resulting small-signal capacitances. For the strong inversion approximation, Ward calculates the source and drain charges as 38 term polynomials, where each term is a product of terminal voltages of as high as fifth order [2.9]. He proposes an approximation for the source and drain charges of

\[
Q_S \approx \frac{3}{5} Q_{\text{Nsat}} + \frac{3}{10} (Q_N - Q_{\text{Nsat}})
\]

(2.42a)

and

\[
Q_D \approx \frac{2}{5} Q_{\text{Nsat}} + \frac{7}{10} (Q_N - Q_{\text{Nsat}})
\]

(2.42b)

where \(Q_{\text{Nsat}}\) is equal to \(Q_N\) evaluated at \(V_D = V_{\text{Dsat}}\). This represents a 60-40 split of the channel charge between source and drain in saturation, and a 50-50 split when \(V_{DS} = 0\).

2.3.2 Other charge partitioning models

Since the first discussion of channel charge partitioning by Ward and Dutton in [2.3], several other authors have also addressed this issue. Many of these models introduce arbitrary partitioning functions. For example, the charge partitioning function introduced by Serhan and Yu [2.17] is
an arbitrary function that resembles the approximation proposed by Ward (2.42). Yang et al [2.16] have also proposed an arbitrary partitioning which is interesting in that it sets $Q_d=0$ in saturation. Their model, therefore, predicts that $C_{dg}$ is zero in saturation, which is in direct conflict with Ward’s model and data.

Taylor et al [2.18] propose an equal partitioning of the channel charge between source and drain, in a similar way to [2.3]. However, they base this assignment on the physical argument that since electrons flow from source to drain, an increase in channel electrons (e.g. when the gate voltage is increasing) will be supplied entirely from the source, while a decrease in channel electrons (e.g. when the gate voltage is decreasing) will occur strictly via the drain. Therefore, they use a 50-50 weighting between source and drain to reflect the average partitioning in time between source and drain charging currents. Unfortunately, this physical argument is incorrect because it ignores the transport current. Positive charge can be supplied from the source if the transport current is dynamically reduced at the source end of the channel, relative to the value expected from static analysis. Similarly, positive charge can be removed from the drain if the transport current is dynamically reduced at the drain end of the channel. For this reason, $Q_d$ need not be zero in saturation, as assumed by Yang et al [2.16]. Although there is a potential barrier
between the end of the channel and the drain, there is also a transport current across this barrier that can be modu-
lated to create the effect of charging currents of electrons across the barrier.

Similar untenable assumptions have also been made by Arreola [2.19] and by Robinson et al [2.20]. Arreola assumed that \( C_{sd} \) and \( C_{ds} \) were zero, while Robinson et al assumed that the channel charging current must come only from the source when only the source voltage varies, and only from the drain when only the drain voltage varies. However, one other author has produced a valid analysis without these kinds of a priori assumptions.

Conilogue with Viswanathan [2.21,2.22] used a perturba-
tion analysis to calculate the source and drain currents from the transport and continuity equations when \( \frac{dQ_n}{dt} \neq 0 \). The first order term in this perturbation corresponds to the low-frequency source and drain capacitances. From this solution, Conilogue identifies the source and drain charges

\[
Q_s = \mu \frac{W}{L_{O}} \int_{V_s}^{V_D} \int_{0}^{\tau} Q_n(\tau) \int_{V_s}^{V_D} Q_n(\phi)^2 d\phi d\tau \quad (2.43a)
\]

and

\[
Q_d = \mu \frac{W}{L_{O}} \int_{V_s}^{V_D} \int_{0}^{\tau} Q_n(\tau) \int_{V_s}^{V_D} Q_n(\phi)^2 d\phi d\tau 
\]

where \( \phi \) and \( \tau \) are dummy variables of integration for \( Q_n(V) \).
By repeated use of

$$Q_n \, dv = \frac{I_o}{\mu W} \, dy \quad ,$$  \hspace{1cm} (2.44)

which comes from (2.25), it is easily seen that (2.43) is identical to (2.41).

Therefore, the only two charge partitioning analyses which proceed from first principles, e.g. Conilogue's and Ward's, can be shown to yield identical partitioning functions for the channel charge. Although this partitioning may produce source and drain charges which are too complicated for use in circuit simulators, any good approximation for the source and drain charges should agree in a general way with this result.

### 2.3.3 A circuit equivalent model for the Ward MOSFET capacitance model

The Ward large-signal capacitance model leads to a small-signal IAM where the 16 capacitance terms are, in general, nonreciprocal. This capacitance matrix is shown in Fig. 2.9. Of the sixteen capacitance terms, only 9 are independent, and therefore a nine element circuit equivalent model can be constructed which embodies the general IAM. The constraints among the 16 capacitance terms are summarized below, where one of the eight constraints is redundant.
\[ C_{gg} = C_{gs} + C_{gd} + C_{gb} = C_{sg} + C_{dg} + C_{bg} \]
\[ C_{ss} = C_{sg} + C_{sd} + C_{sb} = C_{gs} + C_{ds} + C_{bs} \]
\[ C_{dd} = C_{dg} + C_{ds} + C_{db} = C_{gd} + C_{sd} + C_{bd} \]
\[ C_{bb} = C_{bs} + C_{bd} + C_{bg} = C_{sb} + C_{db} + C_{gb} \] (2.45)

The equivalent circuit model will consist of six inter-terminal capacitors and three transcapacitors. However, because of the degeneracy in the IAM, the model will not be unique; many combinations of capacitors and transcapacitors are possible, each choosing a different set of nine fundamental capacitance terms.

\[
\begin{bmatrix}
  C_{gg} & -C_{gs} & -C_{gd} & -C_{gb} \\
  -C_{sg} & C_{ss} & -C_{sd} & -C_{sb} \\
  -C_{dg} & -C_{ds} & C_{dd} & -C_{db} \\
  -C_{bg} & -C_{bs} & -C_{bd} & C_{bb}
\end{bmatrix}
\]

Figure 2.9 - The general four-terminal capacitance matrix.
The model is constructed by choosing 6 off-diagonal terms for the 6 capacitors, and then segregating the IAM into a symmetric part, representing these 6 capacitors, and an asymmetric part, representing the 3 transcapacitors. In this case, the 6 capacitor terms have been chosen to match the capacitors of the Meyer model: $C_{gs}$, $C_{gd}$, $C_{gb}$, $C_{bs}$, $C_{bd}$ from the Meyer model plus $C_{sd}$. (See Fig 2.10.) Three of the terms from the asymmetric matrix are now chosen for the three transcapacitors: in this case, $(C_{dg} - C_{gd})$, $(C_{bd} - C_{bd})$ and $(C_{bg} - C_{gb})$. These three terms were chosen because they each happen to be positive in the Ward model. The asymmetric matrix is then rewritten in terms of the three transcapacitors, using the constraints summarized above as needed (Fig. 2.11). From Fig. 2.11, the placement of the transcapacitors in the circuit equivalent model is clear. The final model is shown in Fig. 2.12.

There are two reasons why this model is of conceptual interest. First, by showing how the four-terminal matrix can be implemented with two-terminal elements, it makes the concept of a four-terminal, nonreciprocal capacitance matrix more concrete. Second, it dramatizes the role of charge partitioning in the device behavior. If the small nonreciprocity of $C_{gb}$ and $C_{bg}$ is ignored, the remaining two transcapacitors, as well as $C_{sd}$, simply redistribute the gate-to-channel and bulk-to-channel capacitances between the drain and source terminals. Also, if $C_{gb} - C_{bg} = 0$, the gate-to-
\[
\begin{pmatrix}
(c_{gs} + c_{gd} + c_{gb}) & -c_{gs} & -c_{gd} & -c_{gb} \\
-c_{gs} & (c_{gs} + c_{sd} + c_{bs}) & -c_{sd} & -c_{bs} \\
-c_{gd} & -c_{sd} & (c_{gd} + c_{sd} + c_{bd}) & -c_{bd} \\
-c_{gb} & -c_{bs} & -c_{bd} & (c_{gb} + c_{bs} + c_{bd})
\end{pmatrix}
\]

\[
\begin{pmatrix}
0 & 0 & 0 & 0 & 0 \\
(c_{gs} - c_{sg}) & (c_{ds} - c_{sd}) & 0 & (c_{bs} - c_{sb}) \\
(c_{gd} - c_{dg}) & (c_{sd} - c_{ds}) & 0 & (c_{bd} - c_{db}) \\
(c_{gb} - c_{bg}) & 0 & 0 & (c_{bg} - c_{gb})
\end{pmatrix}
\]

Figure 2.10 – a.) The symmetric part, and b.) the asymmetric part of the general capacitance matrix.
channel and channel-to-gate capacitances are clearly reciprocal, and the same is true for the bulk-to-channel and channel-to-bulk capacitances.

\[
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 \\
(C_{dg} - C_{gd}) & -(C_{dg} - C_{gd}) & 0 & -(C_{bg} - C_{gb}) \\
+(C_{bg} - C_{gb}) & -(C_{db} - C_{bd}) & + (C_{db} - C_{bd}) \\
-(C_{dg} - C_{gd}) & (C_{dg} - C_{gd}) & 0 & -(C_{db} - C_{bd}) \\
-(C_{bg} - C_{gb}) & 0 & 0 & C_{bg} - C_{gb}
\end{bmatrix}
\]

Figure 2.11 - Transformation of the asymmetric matrix using the constraints of (2.45).
Figure 2.12 - The final circuit equivalent model.
2.4 Non-Quasi-Static Modeling of the MOSFET

Prior to the development of the Meyer model \([2.1]\), there was more interest in non-quasi-static dynamic models for FET's \([2.23-2.35]\) than in quasi-static dynamic models. For both JFET's and MOSFET's, non-quasi-static analysis is equivalent to treating the channel as a nonuniform, nonlinear transmission line. This approach generally requires a linearization of the coupled transport and continuity equations in order to achieve an analytical solution; the solution generally obtained is a set of multi-port admittances which are each an infinite series in \(s\) (complex frequency). As a result, this type of analysis has most frequently been used to construct small-signal, equivalent circuit models for the device.

This section presents a non-quasi-static analysis, similar to earlier work in this field, which is used to further explore the relationship between the Meyer and Ward capacitance models. The Meyer and Ward models are shown to diverge only at high frequencies where the channel capacitances become significant compared to the channel conductances. The non-quasi-static analysis provides a benchmark for the Meyer and Ward models at high frequencies, which highlights the limitations of quasi-static modeling.
2.4.1 Non-quasi-static modeling of the MOSFET

The first use of non-quasi-static analysis for the dynamic modeling of FET's was that of Van Der Ziel and Ero for the junction FET [2.23]. These authors treated the channel as a nonuniform, nonlinear RC transmission line, and derived the wave equation from a gradual channel analysis of the device, similar to Section 2.1 above. Subsequently, other authors have performed similar analyses of both JFET's [2.24, 2.25] and MOSFET's [2.26-2.36]. The resulting wave equations in both cases are nonlinear, and solutions are only possible for linearizations or perturbations around an operating point. Moreover, analytic solutions for the small-signal admittances have only been achieved for the case of zero bulk charge (the three-terminal MOSFET.) Many of these solutions for the MOSFET admittances are further limited to saturation or to a particular two-port configuration (with other terms neglected.) The most complete analysis presented to date is that of Haslett [2.32], which provides a complete, indefinite admittance description over all bias conditions for the three-terminal MOSFET. The analysis presented below [2.36] follows that of Haslett. (It should be noted that the solution presented by Haslett in [2.32] is incorrect, apparently due to an algebraic error.)
The basic transport and continuity equations for an n-channel MOSFET presented in Section 2.1 are reproduced in (2.46) and (2.47) below. Recall that \( V(y) \) is the electron quasi-Fermi potential, \( Q_n(y,t) \) is the channel inversion charge density and \( I(y,t) \) is the channel current as a function of position.

\[
I(y,t) = \mu W Q_n(y,t) \frac{\partial V(y,t)}{\partial y} \tag{2.46}
\]

\[
\frac{\partial I}{\partial y} = -W \frac{\partial Q_n(y,t)}{\partial t} \tag{2.47}
\]

For the three-terminal MOSFET under strong inversion, the mobile channel charge is equal in magnitude to the gate charge and can be approximated by

\[
Q_n(y,t) = -C_{ox}[V_G - V_T - V(y,t)] \tag{2.48}
\]

where \( V_T = V_{FB} + 2\phi_F \) is the threshold voltage. For the three-terminal MOSFET, it is convenient to make a change of variables from the potential along the channel, \( V(y) \), to the potential across the oxide, \( V_x(y) \), which is defined by

\[
V_x(y,t) = [V_G - V_T - V(y,t)] . \tag{2.49}
\]

Using (2.49), the transport and continuity equations can be rewritten as

\[
I(y,t) = \mu W C_{ox} V_x(y,t) \frac{\partial V_x(y,t)}{\partial y} \tag{2.50}
\]
\[ \frac{\partial I(y, t)}{\partial y} = W C_{ox} \frac{\partial V_x(y, t)}{\partial t} . \] (2.51)

Eqs. (2.50) and (2.51) constitute a pair of coupled, nonlinear equations. Unfortunately, a general solution for \( I(y, t) \) in terms of the total terminal voltages is not possible. However, if the terminal voltages are restricted to small variations around an operating point, the incremental current can be calculated as a function of frequency. To facilitate an incrementalization of (2.50) and (2.51), we can rewrite the total channel variables in terms of quiescent and incremental variables as follows:

\[ I(y, t) = I_0(y) + i(y, s)e^{st} \] (2.52a)

\[ V_x(y, t) = V_x(y) + v_x(y, s)e^{st} \] (2.52b)

Similarly, the terminal voltages are assumed to be of the form:

\[ V_G(t) = V_G + v_G(s)e^{st} \] (2.53a)

\[ V_S(t) = V_S + v_S(s)e^{st} \] (2.53b)

\[ V_D(t) = V_D + v_D(s)e^{st} \] (2.53c)

Using this notation, (2.50) and (2.51) can each be rewritten as two equations: one in terms of the quiescent variables alone, and another in terms of the incremental variables.
with the quiescent variables appearing as constants. Equation (2.50) produces

\[
\bar{I}_0(y) = \mu W C_{ox} \bar{V}_X \frac{\partial \bar{V}_X(y)}{\partial y} \quad (2.54a)
\]

\[
i(y,s) = \mu W C_{ox} \frac{\partial [\bar{V}_X(y) v_X(y,s)]}{\partial y} \quad (2.54b)
\]

and Eq. (2.51) produces

\[
\frac{\partial \bar{I}_0(y)}{\partial y} = W C_{ox} \frac{\partial \bar{V}_X(y)}{\partial t} = 0 \quad (2.55a)
\]

\[
\frac{\partial i(y,s)}{\partial y} = s W C_{ox} v_X(y,s) . \quad (2.55b)
\]

From (2.55a) it is again clear that the quiescent transport current, \(I_0(y)\), is constant along the channel, and this constant, \(I_0\), can be found by integrating (2.54a) from source to drain.

\[
I_0 = \frac{\mu C_{ox}}{2} \frac{W}{L} \left[ (V_{GS} - V_T)^2 - (V_{GD} - V_T)^2 \right] \quad (2.56)
\]

Notice that \(V_{GS}\) and \(V_{GD}\) have been used as the independent terminal voltages rather than the more usual \(V_{GS}\) and \(V_{DS}\). This choice is consistent with the source and drain boundary conditions on \(V_X(y)\) and simplifies the algebra which follows. Eq. (2.56) can be simplified to

\[
I_0 = \frac{\mu C_{ox}}{2} \frac{W}{L} (V_{GS} - V_T)^2 (1 - \alpha^2) \quad (2.57)
\]
by the introduction of the variable $\alpha$, which is defined by

$$\alpha = \frac{(V_{GD} - V_T)}{(V_{GS} - V_T)}. \quad (2.58)$$

The coefficient $\alpha$ varies from 1, when the drain-to-source bias is very small, to 0, in saturation.

The next step is to replace $dy$ in (2.54b) and (2.55b) with

$$dy = \frac{\mu W C_{OX}}{I_O} \frac{\mu}{\bar{V}_x} d\bar{V}_x \quad (2.59)$$

to get

$$i(y, s) = \frac{I_O}{\bar{V}_x(y)} \frac{\partial [\bar{V}_x(y) V_x(y, s)]}{\partial \bar{V}_x(y)} \quad (2.60)$$

and

$$\frac{\partial i(y, s)}{\partial \bar{V}_x(y)} = \frac{s \mu (W C_{OX})^2}{I_O} \frac{\bar{V}_x(y) V_x(y, s)}{\bar{V}_x(y) V_x(y, s)} \quad . \quad (2.61)$$

Taking the partial derivative of (2.61) with respect to $\bar{V}_x$, and substituting into (2.60) produces the wave equation

$$\bar{V}_x^2 \frac{\partial^2 i(\bar{V}_x, s)}{\partial \bar{V}_x^2} - \frac{4s}{(1-\alpha)^2} \frac{\bar{V}_x^3}{(V_{GS} - V_T)^3} i(\bar{V}_x, s) = 0 \quad (2.62)$$

where (2.57) has been used for $I_O$, and $\bar{s}$ is the normalized frequency defined by

$$\bar{s} = \frac{s}{\bar{\omega}_o}, \quad (2.63)$$
where
\[ \omega_0 = \frac{\mu (V_{GS} - V_T)}{L^2} = \frac{g_m^*}{C_o} \], \quad (2.64)

\[ g_m^* = g_m + g_d = \mu C_{ox} \frac{W}{L} (V_{GS} - V_T) \], \quad (2.65)

and \( C_o = W L C_{ox} \) is the total oxide capacitance. Eq. (2.62) is a standard Bessel form which leads to a Bessel's function solution, with constants, for \( I(V_X, s) \). The constants can be evaluated using (2.61) and the boundary conditions on \( V_X \) and \( V_X \) at the source and drain. The resulting Indefinite Admittance Matrix (IAM) terms are shown in Table 2.1 truncated to second order in \( \bar{s} \).

2.4.2 Comparison of the Meyer and Ward models to the non-quasi-static solution

The non-quasi-static IAM has the form of a power series matrix over a common denominator series. The numerator matrix is completely symmetric with the exception of the transconductance terms. If we compare the non-quasi-static IAM to the Meyer and Ward capacitances for the zero bulk charge MOSFET, shown in Table 2.2, we see that the first-order terms in the numerator of the non-quasi-static IAM are exact / the Meyer capacitances. So, the Meyer model is the first-order truncation of the numerator of the non-quasi-static solution. The Ward model is a better approximation to the non-quasi-static solution and includes the first-
\[ y_{gg} = \left[ \begin{array}{c}
g_m \bar{s} \frac{2}{3} \frac{1+4a+a^2}{(1+a)^2} + g_m \bar{s}^2 \frac{2}{45} \frac{2+11a+2a^2}{(1+a)^3}\end{array} \right] \cdot \frac{1}{D(\bar{s})} \]

\[ y_{gs} = \left[ \begin{array}{c}
g_m \bar{s} \frac{2}{3} \frac{1+2a}{(1+a)^2} - g_m \bar{s}^2 \frac{2}{45} \frac{2+8a+5a^2}{(1+a)^3}\end{array} \right] \cdot \frac{1}{D(\bar{s})} \]

\[ y_{gd} = \left[ \begin{array}{c}
g_m \bar{s} \frac{2}{3} \frac{a(2+a)}{(1+a)^2} - g_m \bar{s}^2 \frac{2a}{45} \frac{5+8a+2a^2}{(1+a)^4}\end{array} \right] \cdot \frac{1}{D(\bar{s})} \]

\[ y_{sg} = \left[ \begin{array}{c}
-g_m - g_m \bar{s} \frac{2}{3} \frac{1+2a}{(1+a)^2} - g_m \bar{s}^2 \frac{2}{45} \frac{2+8a+5a^2}{(1+a)^3}\end{array} \right] \cdot \frac{1}{D(\bar{s})} \]

\[ y_{ss} = \left[ \begin{array}{c}
(g_m + g_d) + g_m \bar{s} \frac{2}{3} \frac{1+2a}{(1+a)^2} + g_m \bar{s}^2 \frac{2}{45} \frac{2+8a+5a^2}{(1+a)^3}\end{array} \right] \cdot \frac{1}{D(\bar{s})} \]

\[ y_{sd} = \left[ \begin{array}{c}
-g_d
\end{array} \right] \cdot \frac{1}{D(\bar{s})} \]

\[ y_{dg} = \left[ \begin{array}{c}
g_m - g_m \bar{s} \frac{2}{3} \frac{a(2+a)}{(1+a)^2} - g_m \bar{s}^2 \frac{2a}{45} \frac{5+8a+2a^2}{(1+a)^4}\end{array} \right] \cdot \frac{1}{D(\bar{s})} \]

\[ y_{ds} = \left[ \begin{array}{c}
-(g_m + g_d)
\end{array} \right] \cdot \frac{1}{D(\bar{s})} \]

\[ y_{dd} = \left[ \begin{array}{c}
g_d + g_m \bar{s} \frac{2}{3} \frac{a(2+a)}{(1+a)^2} + g_m \bar{s}^2 \frac{2a}{45} \frac{5+8a+2a^2}{(1+a)^4}\end{array} \right] \cdot \frac{1}{D(\bar{s})} \]

\[ D(\bar{s}) = 1 + \bar{s} \frac{4}{15} \frac{1+3a+a^2}{(1+a)} + \bar{s}^2 \frac{1}{45} \frac{1+4a+a^2}{(1+a)} \]

Table 2.1 - The non-quasi-static indefinite admittance matrix terms, truncated to second order.
\[
\begin{align*}
C_{gg} &= C_0 \frac{2}{3} \frac{1 + 4a + a^2}{(1+a)^2}, \quad \text{Meyer} \\
C_{gs} &= C_0 \frac{2}{3} \frac{1 + 2a}{(1+a)^2}, \quad \text{Ward} \\
C_{gd} &= C_0 \frac{2}{3} \frac{a(2 + a)}{(1+a)^2} \\
C_{sg} &= C_0 \frac{2}{3} \frac{1 + 2a}{(1+a)^2} \\
C_{ss} &= C_0 \frac{2}{3} \frac{1 + 2a}{(1+a)^2} \\
C_{sd} &= 0 \\
C_{dg} &= C_0 \frac{2}{3} \frac{a(2 + a)}{(1+a)^2} \\
C_{ds} &= 0 \\
C_{dd} &= C_0 \frac{2}{3} \frac{a(2 + a)}{(1+a)^2} \\
\end{align*}
\]

Table 2.2 - The Meyer and Ward capacitances for the zero bulk charge approximation.

- 91 -
order term of the denominator. This is demonstrated in the following example.

Consider the $Y_{sg}$ term. The non-quasi-static solution truncated to first order is

$$Y_{sg} = \frac{-g_m - g_m^* \frac{2}{3} \frac{1+2a}{(1+a)^2}}{1 + \frac{s}{4} \frac{1+3a+a^2}{15 (1+a)^2}}$$

(2.66)

while the Meyer expression for $Y_{sg}$ is

$$Y_{sg} = -g_m - g_m^* \frac{2}{3} \frac{1+2a}{(1+a)^2}$$

(2.67)

where $g_m^* s = sC_0$. The Ward expression is related to the non-quasi-static solution as follows. If $s$ is small, (2.66) can be approximated as

$$Y_{sg} \approx -g_m - s \left[ g_m^* \frac{2}{3} \frac{1+2a}{(1+a)^2} - g_m \frac{4}{15} \frac{1+3a+a^2}{(1+a)^2} \right]$$

(2.68)

where the second term within the brackets comes from the denominator, using the approximation

$$\frac{1}{1+\tau s} \approx 1-\tau s$$

(2.69)

for small $s$. Comparing (2.68) to Table 2.2, and making use of the expression for $g_m$, it can be shown that the total first-order term above is exactly equal to the Ward capacitance for $C_{sg}$. Thus, the Ward model is the first-order
approximation to the non-quasi-static solution at low frequencies, including the effect of the denominator series.

For the three-terminal MOSFET, therefore, the relationship between the incremental Meyer (M) capacitances and the incremental Ward capacitances (W) can be expressed by

\[ C_{ij}^W = C_{ij}^M - \tau G_{ij} \]  \hspace{1cm} (2.70)

where \( \tau \) is the first-order coefficient of \( D(s) \) from Table 2.1. The interesting feature of (2.70) is the structure of the Ward model that it demonstrates. Although the IAM of the Ward model contains four independent capacitance terms for the three-terminal MOSFET case, these four capacitances can be described in terms of the two Meyer capacitances, plus a single time constant multiplying the two Meyer conductances.

One way to conceptualize the non-quasi-static IAM is to consider the terms in the numerator as being degraded at high frequencies by the common denominator. For example, consider a first-order truncation of the non-quasi-static result. The constant and first-order terms in the numerator are the conductances and capacitances in the absence of high-frequency delays in the channel charge, while the common denominator is the first-order approximation to this degradation. The Meyer model, in this view, represents the low-frequency charging currents, while the Ward model
includes the first-order effect of charging delays on the charge-controlled conductance terms. This interpretation also suggests the origin of the negative source-drain capacitances: these terms are not due to channel charging currents, but instead arise from delays in the conduction currents.

The high-frequency behavior of the Meyer and Ward models are compared in Fig. 2.13 to the non-quasi-static solution, again truncated to second order. The frequency dependence of $Y_{sg}$ is shown for each model, along with experimental data from a long channel (500 μm) device. The bias conditions have been chosen such that the device is saturated, and the admittances are normalized by the DC transconductance. Generally, we see very good agreement between the non-quasi-static theory and the experimental data for frequencies up to twice $\omega_0$. This is particularly significant given that the theory ignores bulk charge which is clearly present for the experimental device. (It would appear that inclusion of bulk charge in the theory would not appreciably change the non-quasi-static solution.) The Ward model tracks the non-quasi-static solution up to about $\omega_0$, but the Meyer model diverges significantly well below $\omega_0$. So, the Ward model appears to be a better high-frequency model than the Meyer model, but the non-quasi-static solution is better still.

- 94 -
Figure 2.13 - Normalized admittance ($Y_{sg}$) versus normalized frequency for the Meyer [1] and Ward [2] models, the second-order truncation of the non-quasi-static solution [3], and experimental data: (a.) magnitude and (b.) angle.
2.4.3 Modeling using the non-quasi-static solution

A first-order truncation of the non-quasi-static solution series is an attractive candidate for an incremental model. The small-signal IAM for the first-order (FO) model can be defined in terms of the Meyer IAM as follows.

\[ y_{ij}^{\text{FO}} = \frac{y_{ij}^{M}}{1 + \tau_s} \tag{2.71} \]

So the first-order model can be described in terms of the simpler Meyer model with the introduction of the time constant \( \tau \). As discussed above, this model agrees with the low-frequency capacitances of the Ward model.

However, the first-order non-quasi-static model has two advantages over the Ward model at high frequencies. First, this model does roll off the gate capacitances at high frequencies, where the quasi-static assumption of the Ward model breaks down. This is demonstrated by the plot of \( C_{gs} \) versus frequency (Fig. 2.14) which compares the Ward (and Meyer) models to the first-order model and actual experimental data. Although the agreement is not as good as for the second-order truncation for \( Y_{sg} \), shown in Fig. 2.13, the first-order model does retain the correct qualitative behavior. This could be important for analog circuits where large-geometry devices are often used for noise or matching reasons, for example, for the input pair of operational
Figure 2.14 - Normalized admittance ($Y_{gs}$) versus normalized frequency for the Meyer [1] and Ward [2] models, the first-order truncation of the non-quasi-static solution [3], and experimental data: (a.) magnitude and (b.) angle.
amplifiers. These large-geometry devices can easily be driven with signals in excess of \( \omega_0 \).

Second, the first-order model does not have the same thermodynamic problem that the Ward model has at high frequencies. The numerator of the first-order truncation is symmetric for all but the DC terms, like the Meyer model, and therefore will not produce unbounded energy at high frequencies. This is doubly the case since the denominator term rolls off all of the small signal admittances for frequencies well above \( \omega_0 \). This property may be of benefit for numerical circuit simulators where artificially fast input signals are sometimes specified by the designer.

An approximate circuit equivalent model for the first-order truncation of the non-quasi-static solution is achieved by adding a bias dependent resistor, \( R_g' \), in series with the gate of the Meyer model, as shown in Fig. 2.15. The incremental voltage at the internal node of this model, \( V_g' \), can be calculated in terms of the three terminal voltages and the simple Meyer admittances as

\[
V_g' = \frac{1}{1 + \tau s} (V_g + R_g Y_{gs} V_s + R_g Y_{gd} V_d) \quad (2.72)
\]

where

\[
\tau = \frac{1}{s} \frac{R_g Y_g}{R_g Y_{gg}} = \frac{R_g C_g}{R_g Y_{gg}} \quad (2.73)
\]

is the time constant found from the non-quasi-static analysis and is defined by (2.71). The three-terminal small-
Figure 2.15 – A circuit equivalent model for the first-order truncation of the non-quasi-static solution in terms of the Meyer admittances.

Signal admittances for this network can be expressed in terms of the Meyer admittances and $\tau$ as

$$y_{ij} = \frac{y_{ij}^M}{1 + \tau s} + \frac{R_g (y_{ij}^M y_{gg}^M - y_{ig}^M y_{gj}^M)}{1 + \tau s}. \quad (2.74)$$

The first term in (2.74) is the desired result and the second term is the error in this equivalent circuit model. It is easy to show that the error is exactly zero if either $i$ or $j$ is the gate terminal; the network exactly models the first-order non-quasi-static IAM for all of the gate admittances (the first row), and for all of the gate excitations (the first column.)
The error, therefore, appears as a simple two-terminal admittance between drain and source, \( y_{sd}' \), and is equal to

\[
y_{sd}' = - \left[ g_d' + (g_m + sC_g) \frac{C_{gd}}{C_{gg}} \right] \frac{\tau s}{1 + \tau s} .
\] (2.75)

The first term within the brackets reflects the fact that the source-drain conductance, \( g_d' \), is unaffected by the addition of the series resistor, while the second term arises from the coupling of the drain voltage to the voltage at \( G' \). For low frequencies, the term \((g_m + sC_g)\) can be approximated by \( g_m \), in which case the second term above can be neglected since \( g_m C_{gd} / C_{gg} \) goes to zero both in saturation \((C_{gd} = 0)\) and for \( V_{DS} = 0 \) \((g_m = 0)\), and therefore should be negligible (compared to \( g_d \)) for bias conditions between these two limits. The error in the direct \( g_d \) term above can be corrected by the addition of a series inductor in the Meyer model, as shown in Fig. 2.16. The series conductance of the conductor-inductor combination is given by

\[
g = \frac{g_d}{1 + sg_d L} ,
\] (2.76)

which gives exactly the desired result if \( L = \tau / g_d \). This inductor corresponds exactly to the negative capacitor between source and drain in the Ward model, except that the inductor has the correct asymptote at very high frequencies.
Figure 2.16 – An improved circuit equivalent model for the first-order truncation of the non-quasi-static solution.
Although the first-order non-quasi-static IAM model and the associated equivalent circuit help to explain the behavior of the device at high-frequencies, these models cannot be applied directly for practical circuit simulation. First of all, the non-quasi-static analysis presented above applies only to the three-terminal, zero bulk charge MOSFET. Unfortunately, the analysis does not lead to a simple solution when the channel charge relationship is generalized to include the effect of the bulk charge. On the basis of the agreement of the three-terminal solution to the real (four-terminal) data in Fig. 2.13, it might not be unreasonable to assume a solution of the same form, and then empirically correct the expression for the denominator time constant to match real device data. This would provide a simpler model than the Ward model which would be free of energy conservation problems, and would be more accurate at high frequencies. However, since this is strictly an IAM description of the device, this model would be only applicable to small-signal analysis.

Unless a similar analysis is discovered which verifies the model for large signals and provides a charge based expression for the denominator time constant, it is not possible to apply the first-order model to transient analysis. It would be possible to introduce the $1/(1+\tau s)$ dependence into the terminal charges and the (charge controlled) transport current through a simple difference equation.
However, this approach would require a charge based description of the Meyer model for the low-frequency source and drain charges. Unfortunately, the Meyer model itself is strictly a small-signal model, and therefore, even an ad hoc large-signal model based on the first-order truncation of the non-quasi-static solution is not possible. As a result, although the non-quasi-static analysis presented here provides some interesting insight into the negative and non-reciprocal capacitances of the Ward model, it is not immediately applicable to computer-based circuit simulation.
CHAPTER 3

Measurement of Small-Geometry
MOS Transistor Capacitances

This chapter discusses a new technique for the measurement of small-geometry MOS transistor capacitances. This technique uses an on-chip operational amplifier in a closed-loop configuration to measure gate and bulk capacitances. The technique has been implemented in a 5 \( \mu \)m CMOS process, and capacitance data for a variety of device geometries have been obtained.

First, the motivation for using an integrated circuit approach to the measurement of small-geometry device capacitances is discussed. Second, the principal general methods for capacitance measurement are reviewed, and the applicability of these methods for on-chip capacitance measurement is examined. Next, the MOS implementation of the closed-loop approach, which has been chosen for this work, is described. Some sample data are presented and the limitations of this approach are discussed.
3.1 Motivation for
an On-Chip Capacitance Measurement Approach

This investigation of on-chip capacitance measurement techniques was motivated by the need for capacitance data from small-geometry devices, and the great difficulty of measuring these small capacitances using conventional techniques. Capacitance data from small-geometry devices are required because of the geometry dependence of the MOS transistor characteristics. These dependencies are frequently described in terms of independent short-channel and narrow-channel effects; however in practice, MOS circuits make use of devices that are minimum-geometry in both length and width. These devices are so small that their capacitances can not be measured using off-the-shelf capacitance instruments and simple wafer probing techniques. As a result, on-chip techniques must be used to measure the capacitance characteristics of individual minimum-geometry devices.

3.1.1 Need for capacitance data of small-geometry devices

As lithographic and pattern definition techniques continue to improve, minimum device geometries are increasingly determined as much by device characteristics as by processing capabilities. Junction depths and lateral oxide encroachment ultimately define minimum channel lengths and
widths respectively for acceptable device performance. Although modern processes are designed to minimize these effects, minimum-geometry devices invariably display some amount of short- and narrow-channel phenomena.

Capacitance data from these small-geometry devices are needed for two reasons. First, these data are needed to support the development of short- and narrow-channel capacitance models. Although short- and narrow-channel effects on the DC device behavior have been investigated extensively, there has been very little work in the area of capacitance modeling for small devices. In spite of some recent theoretical work on small-geometry device capacitances [3.1], most circuit simulators use capacitance models which have been derived for large-geometry devices and then extrapolated to small devices. These models become increasingly inaccurate as integrated circuit technologies evolve, particularly for velocity saturated (short-channel) devices.

Second, small-geometry device capacitance data are needed for device characterization. Even with a good model for these capacitance effects, the devices of a particular process must be fully characterized to obtain the model parameters. Since every technology yields devices which are slightly different, the model parameters must be carefully chosen to achieve accurate circuit simulation for a given technology. It is particularly important to characterize
the actual devices that are used in real circuits, which are frequently minimum-geometry in both width and length. Unfortunately, this is extremely difficult using conventional methods.

3.1.2 Limitations of conventional capacitance measurement techniques

Standard practice for the measurement of small device capacitances involves the use of off-the-shelf capacitance instruments along with wafer probing of the device under test. The best available capacitance meters and bridges provide 0.01 fF resolution for a 2 pF full-scale reading at their most sensitive range. Frequently, however, this resolution is not available at the small signal levels (<100 mV) required for MOS transistor capacitance measurements. Therefore, for high resolution capacitance measurements (of three significant figures,) these meters are typically only capable of measuring devices with a total gate capacitance of greater than or equal to 0.1 pF. However, small-geometry devices can have a total gate capacitance of as little as 1 fF, as for a 1 μm by 1 μm transistor with a 350 Å gate oxide. So, these instruments fall short of the needed resolution by two orders of magnitude, and this is ignoring any linearity or noise problems.
Conventional wafer probing limits the capacitance measurement of small devices by introducing stray and parasitic capacitances. Stray capacitances arise from coupling between the two-point probes as they are brought in contact with the test structure. Parasitic capacitances are associated with the metal pads which must be provided on-chip as targets for these probes, and they generally appear as parasitic terms in parallel with the substrate capacitances of the device. These stray and parasitic capacitances can easily be several tenths of a picofarad. Therefore, using conventional wafer probing, the capacitance instrument must achieve attofarad (aF) resolution on top of a picofarad of offset. Although off-the-shelf instruments have greatly improved in the last several years, it is not expected that these instruments will be improved to the point where accurate capacitance measurements can be made on minimum-geometry, state-of-the-art devices.

As a result of these limitations, the device capacitance data that have been published to date have been taken exclusively from large (100 μm by 100 μm or larger) devices [3.2,5.3]. A 100 μm by 100 μm device corresponds to a total gate capacitance of approximately 5 pF, and it appears that this is about as small a total capacitance as can be conveniently measured using conventional instruments and probing techniques. If only short-channel effects were of interest, a short-channel test device could be used which was
extremely wide (for example 5000 μm wide with a 2 μm channel length) in order to maintain the same total gate area. Similarly, an extremely long but narrow device could be used to investigate narrow-channel effects. In the latter case, however, the $f_T$ of the transistor would be so low that it would be impossible to make a quasi-static capacitance measurement.

The only practical alternative to measuring individual small-geometry devices is to measure many small devices in parallel [3.4]. For example, 2500 2 μm by 2 μm devices would provide the same net gate capacitance as the examples above. Unfortunately, the measured characteristics of the whole population may not be representative of an individual device. For example, using this structure, the sharpness of the measured turn-on characteristics of the gate capacitances near threshold will be affected by the mismatch of $V_T$ among the test devices. Therefore, these capacitance data might be more indicative of the statistical variations of threshold voltage over the population than of the true individual device behavior. This is particularly true for the small-geometry devices of interest, which may have large statistical variations.

Therefore, the only sure way to obtain capacitance data for small-geometry devices is to measure individual devices using on-chip circuits. Since on-chip circuits use ratio-
metric techniques, the resolution in absolute capacitance will scale with the process technology; there is no limit to the ultimate resolution which these techniques can achieve. In addition, on-chip circuits operate in a much smaller stray and parasitic capacitance environment, and these capacitances will also scale with the process technology. So, an on-chip technique which is successful with today's processes should be successful with future technologies as well.
3.2 A Summary of

General Capacitance Measurement Techniques

Before discussing the closed-loop capacitance measurement technique in detail, it is instructive to look at general capacitance measurement techniques to compare their advantages and limitations, and to see how each might be implemented in MOS technologies. These techniques fall into three broad classes: open-loop, closed-loop, and bridge or quasi-bridge techniques. As discussed later, a closed-loop approach has been chosen which combines high accuracy with a simple measurement procedure and a relatively low level of circuit complexity.

3.2.1 Open-loop capacitance measurement techniques

Open-loop capacitance measurement techniques are based upon the simple capacitance divider of Fig. 3.1. The AC gain of the capacitive divider is given by

$$\frac{V_O}{V_{AC}} = \frac{C_{REF}}{C_{REF} + C_t}, \quad (3.1)$$

where $C_t$ is the test capacitor and $C_{REF}$ is a known reference capacitor. The test capacitor can be calculated, in terms of $C_{REF}$, from the measured AC gain using (3.1).
Figure 3.1 - A capacitive voltage divider.

In practice, the AC gain is measured using an off-chip gain-phase instrument. A typical gain-phase meter will have an input capacitance of 10-100 pF, which will appear as a load on the capacitive divider in parallel with $C_t$. In order to measure very small test capacitances, an on-chip buffer must be provided to drive the off-chip gain-phase instrument while introducing as small a load on the capacitive divider as possible. This approach has been used by Iwai and Kohyama to measure small parasitic capacitances [3.5]. A generalized schematic of their technique is shown in Fig. 3.2. The switch must be added to provide a means of calibrating the buffer amplifier and of periodically establishing the DC bias of the floating node.
Figure 3.2 - The open loop capacitance measurement technique of Iwai and Kohyama.

To see how this circuit operates, first consider the case where the amplifier is an ideal unity-gain buffer. When the switch is open, the AC gain of the capacitive divider can be measured from the buffered output \( V_o \). This gain is equal to

\[
\frac{V_o}{V_{AC}} = \frac{C_{REF}}{C_{REF} + C_t + C_p}
\]

(3.2)

where \( C_p \) is the net parasitic capacitance on the floating node, which includes the input capacitance of the amplifier. The sum of \( C_t \) and \( C_p \) can be calculated from (3.2) given the measured AC gain. Two similar structures must be measured.
having a known ratio of test capacitance or parasitic capacitance, in order to separate the parasitic and test capacitance terms. When the switch is closed, the buffer amplifier can be driven directly for calibration purposes, or a DC bias can be stored on the floating node. If there are leakage currents at the floating node, the DC bias must be reapplied periodically to prevent drift of the DC bias across the test capacitor. The operation of the switch must be chosen so as not to interfere with the operation of the gain-phase instrument.

This technique is very simple to implement in MOS technologies. An MOS transistor can be used for the switch, and, for best results, a poly-poly capacitor can be used for the reference capacitor. If two levels of polysilicon are not available, the gate capacitance of an MOSFET with drain and source tied together can be used as the reference capacitor. In the latter case, care must be taken to avoid small-geometry effects and to keep the transistor biased in strong inversion. Both of these structures make excellent reference capacitors, particularly the poly-poly capacitor which can have a voltage coefficient of less than 5 ppm/V. Also, the buffer amplifier for this technique can be very simple. For example, Iwai and Kohyama use a single source follower stage. (It should be noted that the amplifier must have a high input resistance, which is not a problem for MOS amplifiers.)
The greatest problem with this technique is the complexity of the measurement procedure. The AC gain of the buffer amplifier must be calibrated over the complete range of DC bias at the floating node before data can be taken for a given test structure; and two similar test structures must be calibrated and measured in order to calculate the final test capacitor values. In addition, there is a potential for accuracy problems if the two ratioed structures are poorly matched.

Another calibration problem with this technique comes from the charge that is dumped onto the floating node when the switch opens. The switch is implemented by an MOS transistor which stores charge when it is on (the switch is closed.) When the transistor turns off (the switch opens), some of this charge is dumped onto the floating node, and this will perturb the DC bias across the test capacitor. This effect is voltage dependent since the charge stored in the transistor is proportional to $V_{GS} - V_T$, where the source voltage is the DC bias at the floating node and the gate potential is the power supply voltage. The charge dumping effect can easily be large enough to cause significant distortions in the bias voltage for capacitance versus voltage curves. Therefore, the DC response of the amplifier must also be calibrated so that the effect of the charge dumping can be measured and the bias sweep pre-distorted to compensate for it. This complicates the measurement proce-
duce even more, and increases the likelihood of matching and calibration errors. The closed-loop technique described next avoids these problems, providing a simpler and more accurate measurement.

3.2.2 Closed-loop capacitance measurement techniques

The closed-loop capacitance measurement technique considered here is taken from the common electrometer circuit of Fig. 3.3a. This circuit is a transresistance amplifier; the output voltage is equal to the input current scaled by the feedback resistor. For capacitive test currents, the coulombmeter circuit of Fig. 3.3b is more appropriate. This circuit measures net charge rather than current. In particular, if $i_t$ is a small-signal, sinusoidal current of the form $sC_tV_t$, where $V_t$ is a small-signal AC voltage and $C_t$ is some effective capacitance, the output is frequency independent and

$$
\frac{V_o}{V_t} = -\frac{C_t}{C_F}
$$

(3.3)

Since the inverting terminal of the opamp is a virtual ground, we can measure such a test capacitor with the simple connection of Fig. 3.3c.

This circuit has two principal attractions. First, it is insensitive to parasitic capacitances on the virtual
Figure 3.3  -  a.) The electrometer circuit, b.) the coulombmeter circuit, and c.) the coulombmeter circuit used to measure capacitance.
ground (or summing) node, and second, it is a ratiometric technique which depends only upon the feedback (or reference) capacitor $C_F$. Essentially, feedback has been added to the open-loop circuit of Fig. 3.2, eliminating the need to know the stray capacitance by ensuring that all of the test current is forced through the feedback capacitor. Therefore, the test capacitor can be determined, relative to $C_F$, by a single gain measurement without any calibration. The closed-loop circuit can also be easily implemented in MOS technologies since NMOS and CMOS operational amplifiers of sufficient quality can readily be designed using established techniques [3.6,3.7].

The insensitivity of this circuit to stray capacitances on the summing node can be seen by considering the block diagrams of Fig. 3.4. The transfer function for this feedback network is approximately $-C_t/C_F$ as long as the loop transmission is much larger than one. The loop transmission, L.T., is given by

$$- \text{L.T.} = \frac{C_F}{C_t + C_F + C_p} \ a(s) \quad (3.4)$$

where $C_p$ is the net stray capacitance on the summing node including the input capacitance of the operational amplifier. (It is assumed that the amplifier has an infinite input resistance.) The effect of the stray capacitance is to reduce the loop transmission, which has no effect on the
Figure 3.4 - Block diagrams for the closed-loop coulombmeter circuit: a.) direct form, and b.) reduced form.
overall transfer function as long as the amplifier gain at
the signal frequency is large enough.

From Fig. 3.3c it is clear that some kind of DC feed-
back around the amplifier is required in order to stabilize
the DC bias at the summing node. This is achieved with a
feedback switch similar to that of the open-loop circuit of
Fig. 3.2. The details of the feedback switch, and the
associated charge dumping, are deferred to Section 3.3 where
the closed-loop coulombmeter technique is discussed in more
detail.

3.2.3 Bridge and quasi-bridge capacitance measurement
techniques

The basic concept of a variable element bridge is
demonstrated by the simplified Wheatstone bridge of
Fig. 3.5. If \( R_x \) is a continuously variable and continuously
calibrated potentiometer, the test resistor, \( R_t \), can be
found by varying \( R_x \) until the current across the bridge goes
to zero. At this point the bridge is balanced and the value
of \( R_t \) is equal to the selected value of \( R_x \). In practice, \( R_x \)
may be only discretely variable, in which case the bridge
current can be used to interpolate between the two settings
of \( R_x \) which come closest to balancing the bridge.

The Wheatstone bridge can be used equally well to
measure capacitors, in which case the applied voltage and
the bridge current will be small-signal sinusoidal quantities. In this case, the detection of a balanced condition in the bridge is equivalent to detecting zero energy in the bridge current at the applied frequency. The advantage of bridge methods is that this detection can be made using nonlinear techniques such as synchronous detection with analog multipliers or saturating gain stages.

The floating electrometer of the Wheatstone bridge can be replaced by a single-ended voltage meter and an inverter in the modified bridge of Fig. 3.6. As before, the bridge is balanced when a null is detected at the output voltage, which occurs when the currents in the two capacitors cancel at the common node. For on-chip capacitance measurements,
however, a technique is required which does not call for a continuously variable element. The quasi-bridge of Fig. 3.7 solves this problem by replacing the variable capacitive element with a variable gain (or attenuation) stage.

This circuit works the same way as the modified bridge above except that the variable capacitor is replaced by a fixed reference capacitor, and the current nulling is achieved by varying the amplitude of the driving voltage on the reference capacitor. The test capacitor is therefore determined by the reference capacitor times the gain that is required to balance the bridge. If the gain stage is not calibrated, the gain can be measured using a gain-phase instrument.
The attractive feature of this circuit for on-chip measurements is that very little of the bridge needs to be on-chip. In particular, only the reference capacitor and a buffer amplifier are required on-chip, as shown in Fig. 3.8. This circuit is essentially identical to the open-loop technique above [3.5]; the only difference is that here both the test capacitor and the reference capacitor are driven with AC voltages. For the quasi-bridge, it is only necessary to detect a null at the floating node, and therefore, the parasitic capacitance and the actual gain of the buffer amplifier have no effect on the final measured capacitance. However, the parasitic capacitance reduces the ultimate sensitivity of the bridge by reducing the gain from the mismatch in currents to the output voltage.
Figure 3.8 - The on-chip circuitry required for the quasi-bridge technique.

Although the quasi-bridge needs very little on-chip circuitry, the off-chip circuitry needed to close the loop is quite extensive. As a result, the closed-loop technique was chosen since it provides good performance with a commercially available instrument doing most of the work. The gain-phase instrument used with the closed-loop coulombmeter in this study was a Hewlett-Packard HP-4192A, which actually uses the quasi-bridge method described above. In this way, the noise benefits and other advantages of synchronous detection in bridge techniques can be realized with the closed-loop technique as well.
3.3 The Closed-Loop Coulombmeter Technique

The closed-loop technique introduced in general terms in the previous section is discussed in detail below. The most important consideration is the DC feedback switch, and the effect that the charge dump of this switch has on the gain-phase instrument. The CMOS implementation of the technique is described, and some sample data are presented which demonstrate the capabilities and limitations of this approach.

3.3.1 Details of the closed-loop coulombmeter technique

The closed-loop technique discussed in section 3.2.2 can be used to measure two-terminal parasitic capacitors or multi-terminal device capacitances. The detailed schematic of the coulombmeter technique, as used to measure gate capacitance, is shown in Fig. 3.9a. Any of the three gate capacitances (C_{gs}, C_{gd}, or C_{gb}) can be measured using this circuit by applying an AC voltage to the corresponding terminal (source, drain, or bulk). For example, if an AC voltage is applied to the source terminal, the magnitude of the current at the gate, which is sensed across the feedback capacitor, is \omega V_{AC} C_{gs}. The AC voltage at the output, in this case, is C_{gs}/C_F times the applied AC voltage, with an inversion. A gain-phase instrument is used to measure the gain from the driven terminal to the output,
Figure 3.9 - Detailed schematic of a.) the gate coulombmeter circuit, and b.) the bulk coulombmeter circuit.
producing a measure of the driven capacitance relative to \( C_F \). (If the measurement frequency is low enough to ensure a large loop transmission, the measured phase will be very close to 180 degrees.) Several devices can be measured using the same amplifier and reference capacitor as long as only one device at a time is driven with an AC voltage.

The source, drain, and bulk terminals are brought directly off-chip, so off-chip circuits can be used to apply the DC biases and the AC signal directly to these terminals. This also makes possible the measurement of the static transport current. The DC bias at the gate is controlled by the voltage applied to the non-inverting input of the operational amplifier, which is also brought directly off-chip. The substrate terminal of the device must be available to measure \( C_{gb} \) or to measure the three bulk capacitances \( (C_{bs}, C_{bd}, \text{ and } C_{bg}) \) as shown in Fig. 3.9b. A CMOS implementation of the closed-loop coulombmeter was chosen for this reason, using test devices in individual wells.

The feedback switch is required to periodically establish the DC bias of the summing node, as in the open-loop technique of Iwai and Kohyama (Fig. 3.2). Without the switch, small leakage currents at the summing node would be integrated by the feedback capacitor until the amplifier was driven out of its linear range. The operation of the switch is as follows. When the switch is closed, the output is
driven to the gate (or bulk) bias voltage (applied at the non-inverting input of the amplifier) and \( C_F \) is discharged. When the switch opens, the output stays near the summing node potential and the AC signal path is enabled. The reset pulse can be very brief (1 to 5 \( \mu \)s depending on the speed of the operational amplifier,) while the reset period can be very long (1-100 ms depending on leakage currents and the size of the feedback capacitor.) The resulting output waveform is shown in Fig. 3.10, where the lower trace is a expansion of the upper trace during the reset pulse. This reset scheme has proven to be both simpler and more accurate than the DC feedback scheme first proposed for the coulomb-meter technique in [3.8].

Figure 3.10 - Oscilloscope photograph of the gate coulomb-meter circuit output.
Charge dumped from the switch as it opens is absorbed by the feedback capacitor, producing a step discontinuity in the output. An extreme example of this is shown in Fig. 3.11a. The magnitude of this voltage step for the closed-loop technique is much larger than for the open-loop technique of Iwai and Kohyama [3.5], because the effective capacitance (in this case just $C_F$) is much smaller. However, for the coulombmeter circuit, this effect does not disturb the bias on the device under test, so no calibration is necessary to correct for it.

In practice, the loss of signal during the reset interval and the output step caused by charge dumping are problems only to the degree to which they disturb the gain-phase instrument, which generally must operate over many reset periods. The most important consideration is that the instrument remain phase locked to the AC signal. A secondary consideration is noise in the gain-phase measurement produced by interference between the AC signal and the reset clock. This can be minimized by keeping the reset repetition rate as far below the signal frequency as possible, and by keeping the two frequencies harmonically unrelated. The effect of charge dumping can also be reduced by using the circuit of Fig. 3.12. If the coulombmeter output is high-pass filtered, it can then be clamped to ground during the reset pulse using an analog switch. The high-pass filter is required to remove the varying DC bias of the coulombmeter.
Figure 3.11 - a.) An oscilloscope photograph of the gate coulombmeter circuit output with large charge dumping. b.) An oscilloscope photograph of the output of the charge dumping suppression circuit of Fig. 3.12 with a.) as the input.
circuit output before the clamp is applied. Figure 3.11b shows the result of using this circuit on the waveform of Fig. 3.11a. Although the large (5 μs wide) pulse is removed, the output still has a residual bump due to the action of the high-pass filter. Nevertheless, this gentler waveform produced much better results than the waveform of Fig. 3.11a when using the HP-4192A as the gain-phase meter.

This source of noise could be eliminated completely using a triggered gain-phase instrument with a short acquisition time. A sample of the AC waveform could be acquired after the disturbances produced during a given reset pulse have settled and before the next reset pulse arrives.
approach was not used here since such an instrument is not commercially available. However, the quality of the data from the closed-loop coulombmeter technique could be greatly improved if such an instrument was developed.

3.3.2 The CMOS implementation and experimental results

The closed-loop coulombmeter technique was implemented in a 5 \( \mu \text{m} \), p-well CMOS process. N-channel transistors in individual wells were used as the test devices to allow the measurement of substrate capacitances. Gate and bulk capacitance structures were included with test device geometries of 100/100, 20/20, 20/4, 4/20 and 4/4 \( W(\mu \text{m})/L(\mu \text{m}) \). (The 4 \( \mu \text{m} \) geometries are smaller than the design rules allow for this process.) Process and device characteristics corresponding to these n-channel devices are listed in Table 3.1. Notice that the length and width reductions for these devices are severe, and that the body effect factor (\( \gamma \)) is quite high. In particular, the effective channel length of a 4 \( \mu \text{m} \) drawn length device is only 1.4 \( \mu \text{m} \) (which is comparable to the junction depth of 1.6 \( \mu \text{m} \)), and the effective channel width of a 4 \( \mu \text{m} \) drawn width device is only 1.1 \( \mu \text{m} \). Not surprisingly, as will be seen later, the small geometry devices exhibit strong short- and narrow-channel effects, including velocity-limited saturation.
Table 3.1

The device and process characteristics.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gate dielectric thickness (t_{OX})</td>
<td>700 Å</td>
</tr>
<tr>
<td>Poly-Poly dielectric thickness</td>
<td>1400 Å</td>
</tr>
<tr>
<td>Mobility, low field (μ)</td>
<td>725 cm²/Vsec</td>
</tr>
<tr>
<td>Threshold Voltage (V_T)</td>
<td>0.7 V</td>
</tr>
<tr>
<td>Body Effect Factor (γ)</td>
<td>1.04 V⁻¹</td>
</tr>
<tr>
<td>P-well Doping, background</td>
<td>6.5 x 10¹⁵ cm⁻³</td>
</tr>
<tr>
<td>P-well Doping, from γ at V_{SB} = 0v</td>
<td>8.3 x 10¹⁵ cm⁻³</td>
</tr>
<tr>
<td>Width Reduction (W_{corr})</td>
<td>2.9 μm</td>
</tr>
<tr>
<td>Length Reduction (L_{corr})</td>
<td>2.6 μm</td>
</tr>
<tr>
<td>Junction Depth (x_j)</td>
<td>1.6 μm</td>
</tr>
</tbody>
</table>

design with internal compensation and a source follower output stage. This amplifier is a general purpose design that could be used for a wide range of CMOS analog circuits. It differs from typical switched capacitor filter amplifiers only in that it is designed to drive a large off-chip capacitive load. The performance of the operational amplifier is summarized in Table 3.2, and a schematic of the amplifier is shown in Fig. 3.13. In this design, a bias network has been developed which makes use of the vertical NPN bipolar transistor formed by an n⁺-diffusion, the p-well, and the n-substrate. There are two special design considerations for this application. First, the voltage swing of the amplifier
Table 3.2

Operational Amplifier Characteristics

<table>
<thead>
<tr>
<th>Parameter (1)</th>
<th>Simulated</th>
<th>Measured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max. Power Supply</td>
<td>-</td>
<td>&gt; 15 V</td>
</tr>
<tr>
<td>Power Dissipation</td>
<td>3.7</td>
<td>4.5 mW</td>
</tr>
<tr>
<td>Open-Loop DC Gain</td>
<td>102</td>
<td>87 dB</td>
</tr>
<tr>
<td>Gain-Bandwidth Product</td>
<td>1.7</td>
<td>2.2 MHz</td>
</tr>
<tr>
<td>Slew Rate</td>
<td>3.0</td>
<td>3.3 V/μs</td>
</tr>
<tr>
<td>Max. Capacitive Load (2)</td>
<td>100</td>
<td>65 pF</td>
</tr>
<tr>
<td>CMRR</td>
<td>102</td>
<td>88 dB</td>
</tr>
<tr>
<td>PSRR (from V_{DD})</td>
<td>112</td>
<td>88 dB</td>
</tr>
<tr>
<td>PSRR (from V_{SS})</td>
<td>113</td>
<td>97 dB</td>
</tr>
<tr>
<td>Input Range</td>
<td>-4.3 to +4.3</td>
<td>V</td>
</tr>
<tr>
<td>Output Range</td>
<td>-4.8 to +3.6</td>
<td>V</td>
</tr>
</tbody>
</table>

(1) All data are for ±5 V power supplies except Max. Power Supply.

(2) Maximum capacitive load for 45° phase margin.
Figure 3.13 - Schematic of the CMOS operational amplifier.
should be as large as possible, and second, the input transistors should not be excessively large.

With careful design using large-geometry devices, the maximum supply voltage of an operational amplifier is limited only by the junction breakdown voltages. Therefore, although the input and output voltage ranges of the amplifier will be 2 to 3 volts less than the maximum power supply voltage, the amplifier does not greatly limit the range of biases that can be applied to the device under test. In this case, the operational amplifier was powered by a 15 volt supply, and the device bias voltages were kept between 2 volts and 12 volts.

Although the closed-loop coulombmeter technique is insensitive to parasitic capacitance on the summing node, the suppression of the parasitic capacitance is not perfect since the amplifier gain at the signal frequency is not infinite. Therefore, there is an upper limit on the size of the input devices of the operational amplifier. In this case, good results were obtained using 40 \( \mu \text{m} \) by 10 \( \mu \text{m} \) p-channel transistors. For a smaller design rule technology with smaller test devices, either the area of the input transistors or the signal frequency would have to be reduced. Unfortunately, as the signal frequency is reduced the charge dumping phenomenon produces more noise in the gain-phase meter. Therefore, in all likelihood, it will be
necessary to scale the input devices with the devices under test. Although the low-frequency (1/f) noise of the amplifier will increase for smaller input devices, this should not be a problem for the coulombmeter circuit since the device noise is insignificant relative to the noise introduced by charge dumping.

The feedback switch (Fig. 3.14) was implemented as a p-channel/n-channel pair of minimum-geometry (5 μm by 5 μm) transistors. A p-n pair was chosen for good dynamic range, and minimum-geometry devices were used to minimize the charge dumping. (Note that minimum channel length devices can not be used as switches if these devices have large

Figure 3.14 - A complementary p-n switch.
drain-to-source leakage currents for gate voltages below $V_T$.) Some care must be taken to ensure the charge dumped from the p-channel device tends to cancel the charge dumped by the n-channel device; however, this cancellation is only effective near the middle of the power supply range where the 'on' voltages of both devices are approximately the same.

The feedback capacitors were poly-poly capacitors, and the capacitor areas were chosen to roughly match the total gate area of the device under test. (The capacitor sizes are listed in Table 3.3.) Smaller feedback capacitors, which produce higher gain, were used for the bulk coulomb-meters since the substrate capacitances tend to be less than the gate capacitances. No attempt was made to determine the absolute value of the feedback (reference) capacitors. Instead, the actual gate and bulk capacitance data were normalized to the total measured intrinsic or extrinsic capacitance. This normalization provides the most meaningful basis of comparison for data from different device geometries. The details of this normalization are included along with the complete data in Appendix A.

In retrospect, the choice of the capacitor sizes was unfortunate. The gate and bulk reference capacitors should have been the same size to facilitate comparison of the gate and bulk data, and all of the capacitors should have been
Table 3.3
Feedback Capacitor Dimensions, W(μm)/L(μm)

<table>
<thead>
<tr>
<th>Test Device</th>
<th>Gate Coulombmeter Feedback Capacitor</th>
<th>Bulk Coulombmeter Feedback Capacitor</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 x 100</td>
<td>100 x 100</td>
<td>100 x 25</td>
</tr>
<tr>
<td>20 x 20</td>
<td>20 x 20</td>
<td>20 x 5</td>
</tr>
<tr>
<td>20 x 4</td>
<td>20 x 5</td>
<td>5 x 5</td>
</tr>
<tr>
<td>4 x 20</td>
<td>20 x 5</td>
<td>5 x 5</td>
</tr>
<tr>
<td>4 x 4</td>
<td>5 x 5</td>
<td>5 x 5</td>
</tr>
</tbody>
</table>

10 μm by 10 μm or greater to avoid small-geometry effects in the poly-poly capacitor values. The very small feedback capacitors also aggravated the charge dumping problem and degraded the loop transmission for the coulombmeter circuits with the smallest test devices.

Data were taken with the help of a microcomputer-based data acquisition system. A simple diagram of the setup is shown in Fig. 3.15. Programmable voltage sources (digital-to-analog converters) were used to set the bias voltages for the test device, and simple (off-chip) operational amplifier circuits were used to add the AC test signal to the DC bias for the driven terminal. The actual signal on the device terminal was used as the reference for the gain-phase
Figure 3.15 – Diagram of the measurement system.
measurements, so that the measured gain was not affected by scale factor errors in the summing circuits. A slight gain error was introduced by the high-pass filter of the charge dump suppression circuit; however, this error was common to all of the data and was eliminated in the normalization procedure. The pulse repetition rate was 400 Hz and the AC signal was approximately 40 kHz (actually 40.123 kHz), except for the very smallest devices for which 10 kHz (actually 10.123 kHz) was used to improve the loop transmission. The amplitude of the AC signal was 100 mV peak to peak, which is equal to the step increment of the sweep voltage ($V_{GS}$). Some sample data are discussed below.

Figures 3.16-3.18 show the intrinsic gate-to-drain capacitances ($C_{gd}$) for three different device geometries: 20/20, 20/4, and 4/4 $W(\mu m)/L(\mu m)$. The solid curves shown are constructed by drawing straight lines between actual data points. The data are normalized to $C_{oi}$, which is the intrinsic gate capacitance as extracted from measured data. Conceptually, $C_{oi}$ is that part of the total oxide capacitance which is defined by the intrinsic (or effective) channel length and the intrinsic (or effective) channel width of the transistor. With this normalization, $C_{gd}$ is equal to 0.5 for $V_{DS}=0$ and $V_{GS} \gg V_T$. (For this bias condition, the gate-to-channel capacitance is shared equally between drain and source, and the intrinsic gate-to-bulk capacitance is zero.)
Figure 3.16 - Experimental intrinsic $C_{gd}$ of a 20 $\mu$m by 20 $\mu$m (W/L) device, normalized by $C_{oi}$.

Figure 3.17 - Experimental intrinsic $C_{gd}$ of a 20 $\mu$m by 4 $\mu$m (W/L) device, normalized by $C_{oi}$.
Figure 3.18 - Experimental intrinsic $C_{gd}$ of a 4 $\mu$m by 4 $\mu$m (W/L) device, normalized by $C_{oi}$.

Comparing Figs. 3.16 and 3.17, for which the channel width is unchanged, the short-channel effects on $C_{gd}$ can be observed. Notice that the transition from the saturation region to the linear region as $V_{GS}$ increases has a noticeably different shape for the 4 $\mu$m long device than for the 20 $\mu$m long device. In particular, for the 4 $\mu$m long device there is a significant increase in gate-to-drain capacitance before the device comes out of saturation.

Figure 3.18 shows $C_{gd}$ for the 4 $\mu$m by 4 $\mu$m device. The total intrinsic gate capacitance for this device is only about 1.0 ff due to the severe width and length reductions. Despite the increased noise, the capacitance resolution is
on the order of 2% of $C_{o1}$, or about 20 aF. The noise in this case is attributable to substantial steps in the output voltage caused by the large effect of charge dumping on the very small feedback capacitor. These steps represent substantial energy at the reset clocking rate, which produces noise in the gain-phase instrument as noted above.

The noise in this measurement could have been reduced by a factor of 5-10 if a charge-compensated switch had been used [3.9]. This technique is shown for a single n-channel switch in Fig. 3.19. The switch is formed by two identical n-channel transistors, M1A and M1B. Two more identical devices, M2 and M3, are placed on either side of the switch with source and drain tied together and are driven by a complementary clock. If the clock switches relatively quickly, it is reasonable to assume that the charge dumped by M1A and M1B as these devices turn off will be evenly divided between the two nodes on either side of the switch. The dumped charge will therefore be canceled by the charge absorbed by M1 and M3 as these devices turn on. In contrast to the n-p complementary switch described above (Fig. 2.14), this charge cancellation will work at any voltage level because the compensating devices always see exactly the same 'on' voltage as the switching devices. Unfortunately, the compensated switch works best when the two nodes on either side of the switch have approximately the same impedance to ground. In this case, one side of the
switch is a small parasitic capacitance and the other side is the output of an operational amplifier. So, the cancellation will not be perfect, but it will be much better than the results obtained here. Notice that two of these compensated switches, one n-channel and one p-channel, are needed in order to retain good switching near both power supplies.

The accuracy of the coulombmeter circuit, and the validity of the data presented here, was verified by measuring the test circuit of Fig. 3.20. In this case, the test capacitor is a poly-poly capacitor of identical construction to the feedback capacitor. The transfer function for this circuit should be exactly equal to one for any DC bias at the input or at the summing node \( V_X \). The result for the
Figure 3.20 - Test circuit for the CMOS coulombmeter implementation.

Case of a 20 \( \mu m \) by 20 \( \mu m \) capacitor is shown in Fig. 3.21, where \( V_X \) varies from 2.5 v to 10.5 v along the x-axis. (The data have been normalized to one; the actual gain error, including the effect of the charge dump suppression circuit, was only 0.1 dB.) The three curves represent three different biases at the input: 2.5 v, 3.5 v, and 4.5 v. The low voltage coefficient of the poly-poly capacitor (<20 ppm/V) is demonstrated by the fact that the three curves lie virtually on top of one another. As the summing node voltage varies over 8 v, the measured gain varies by 0.2% for a scale factor voltage coefficient of approximately 200 ppm/V. This error comes from slight variations in the amplifier gain and voltage coefficients on the summing node parasitic.
Figure 3.21 - Data from the coulombmeter test circuit.

capacitance. These variations are not perfectly eliminated because of the finite gain of the operational amplifier. Better results would have been obtained using a lower signal frequency, however, this would have increased the noise from charge dumping as discussed above. A similar structure with a 20 \( \mu \text{m} \) by 5 \( \mu \text{m} \) capacitor produced a total variation of 0.5\%. So, the 20 \( \mu \text{m} \) by 20 \( \mu \text{m} \) and 20 \( \mu \text{m} \) by 4 \( \mu \text{m} \) device data presented here should have negligible errors due to the measurement technique. The error for the 4 \( \mu \text{m} \) by 4 \( \mu \text{m} \) device data, however, is probably five times larger than the 20 \( \mu \text{m} \) by 5 \( \mu \text{m} \) error.

Figures 3.22 and 3.23 show the intrinsic bulk-to-drain capacitances (\( C_{bd} \)) for a 20 \( \mu \text{m} \) by 20 \( \mu \text{m} \) device and a 4 \( \mu \text{m} \) by
Figure 3.22 - Experimental intrinsic $C_{bd}$ of a 20 $\mu$m by 20 $\mu$m (W/L) device, normalized by $C_{oi}$.

Figure 3.23 - Experimental intrinsic $C_{bd}$ of a 4 $\mu$m by 20 $\mu$m (W/L) device, normalized by $C_{oi}$.
20 \mu m device respectively. The data are again normalized to \( C_{oi} \). For this process, the intrinsic bulk-to-drain capacitances are significant compared to the corresponding gate-to-drain capacitances. For example, the \( V_{DS}=0v \) curve of \( C_{bd} \) in Fig. 3.22 rises to \( 0.3C_{oi} \), which is 70 percent of the corresponding \( C_{gd} \) curve in Fig. 3.16. This ratio of \( C_{bd} \) to \( C_{gd} \) is close to one because of the large body effect of the n-channel devices in the relatively heavily doped p-well. Comparing the \( V_{DS}=0v \) curves of Figs. 3.22 and 3.23, the narrow-channel effects on \( C_{bd} \) can be observed. For the 4 \mu m wide device (Fig. 3.23), \( C_{bd} \) continues to increase as \( V_{GS}-V_T \) increases rather than approaching a constant, as does \( C_{bd} \) for the 20 \mu m wide device (Fig. 3.22). This is a result of the increased effective width of the channel as the larger gate bias inverts more of the area under the encroaching field oxide. This effect, as well as the short-channel effects noted above, will be discussed in Chapter 4.

Similar bulk capacitance curves were generated for the other test device geometries; however, it proved to be impossible to measure the bulk capacitances of 4 \mu m channel length devices at higher drain voltages. For these devices, a drain-to-source bias of 4 V or more was enough to induce substantial bulk hot-electron currents. These currents were so large that the output of the coulombmeter circuit was driven into saturation in less than 1 ms. The coulombmeter output under these conditions is shown in Fig. 3.24. Ironi-
Figure 3.24 - Oscilloscope photograph of the bulk coulombmeter circuit output with excessive bulk hot-electron currents.

Consequently, although this phenomenon is a problem when measuring capacitances, it may prove to be a useful way of measuring hot-electron currents in very small devices.

3.3.3 Limitations of the closed-loop coulombmeter technique

The above described problem points out the major limitation of the closed-loop coulombmeter technique, which is that the technique can only be used to measure pure capacitances, or at least capacitances with only modest leakage currents. This is generally not a problem for parasitic elements or for MOSFET gate capacitances. However, as discussed in the previous section, substrate hot-electron
currents in short-channel devices can be large enough to preclude measuring the bulk capacitances. (This may also be true for gate capacitances in extreme cases, e.g., very short channels with very large drain-to-source bias.)

This also means that the coulombmeter technique can not be used to measure the source and drain capacitances because of the large transport current at the source and drain terminals. Conceptually, the source and drain admittances can be measured using the electrometer circuit of Fig. 3.3a, which uses a resistor as the feedback element. The source and drain capacitances, which are the imaginary parts of the source and drain admittances, can be calculated from the gain and phase of the measured admittances. Unfortunately, the imaginary parts of the source and drain admittances are very small compared to the real parts for frequencies much less than $f_t$. Therefore the phase will be too small to accurately measure the source and drain capacitances of short-channel devices unless the measurement frequency is in the 10's or 100's of MHz. Since the measurement frequency must be much less than the gain-bandwidth product of the operational amplifier ($\approx$1-5 MHz for typical designs), it is impossible to measure these capacitances using the closed-loop technique. The quasi-bridge technique, discussed in Section 3.2, is a good candidate for an on-chip technique for measuring the source and drain capacitances since it does not have the same frequency limitation as the closed-
loop technique. (A general admittance can be measured with the quasi-bridge technique by varying the phase as well as the gain of the reference capacitor's AC voltage.)

A second limitation of the closed-loop coulombmeter technique is that it requires a reasonably stable and well characterized process to support the fabrication of a unity-gain stable operational amplifier. This is not a problem for fundamental device modeling investigations (our principal concern) or for on-going process monitoring; however, it may be a problem in the early stages of device or process design. In practice, this may not be that severe a limitation. Operational amplifiers can be successfully designed in any MOS process, analog or digital, with or without extra features such as poly-poly capacitors or precision resistors. Moreover, a conservative amplifier design with off-chip biasing would probably be successful even in a completely uncharacterized process. Nevertheless, some basic on-chip circuit capability must be provided to implement the closed-loop coulombmeter technique.
CHAPTER 4

Modeling of Small-Geometry MOS Transistor Capacitances

This chapter examines the theory and modeling of small-geometry capacitance effects. Short- and narrow-channel effects are discussed separately, drawing on the experimental capacitance data for 20 \( \mu \text{m} \) by 4 \( \mu \text{m} \) (W/L) and 4 \( \mu \text{m} \) by 20 \( \mu \text{m} \) devices respectively. In each case, the physical mechanisms which produce the short- or narrow-channel effects are qualitatively discussed, and the existing modeling in the literature of the small-geometry static current characteristics is reviewed. Next, the experimental capacitance data for the small-geometry devices is carefully compared to data from larger device, and the small-geometry capacitance effects are identified. Finally, simple models for the small-geometry capacitances are described which accurately predict the observed small-geometry capacitance effects. Narrow-channel effects are discussed first.

(This chapter makes extensive use of the experimental capacitance data presented in Appendix A. It is assumed that the reader is familiar with this material, including the description of the normalization calculations.)
4.1 Narrow-Channel Capacitance Effects

Narrow-channel effects can result from a variety of physical mechanisms, depending upon the physical structure of the device. For practical devices, narrow-channel effects may be attributable to the lateral encroachment of the field oxide, the lateral encroachment of the field implant, or some combination of the two. Regardless of the mechanism, the principal narrow-channel effects are an increase in the threshold voltage and an increase in the effective channel width for increasing gate bias. The latter effect will be shown to have more impact on the device capacitances. These phenomena have been studied by several authors, who have investigated a variety of physical structures [4.1-4.11]. This work is reviewed below in conjunction with a study of the experimental capacitance data. In addition, a simple model for the observed increase in the device capacitances with increasing gate voltage is proposed.

4.1.1 Theoretical basis for narrow-channel effects

Narrow-channel effects were first discussed qualitatively in the literature by Kroell and Ackermann in 1975 [4.1]. The primary interest of later researchers has been the shift of threshold voltage for narrow devices [4.2-4.7]. However, some authors have also addressed the phenomenon of
increasing effective channel width for increasing gate voltage [4.8-4.11]. In particular, Kotecha and Beilstein [4.8] have attempted to calculate the effect of this phenomenon on device capacitances. The sources of narrow-channel effects for a variety of device structures are reviewed below, with particular attention given to the tapered oxide structure produced by the lateral oxidation of the field oxide.

The simplest structure which gives rise to narrow-channel effects is a gate of finite width over a uniform oxide. This simple structure was discussed by Jeppson in 1975 [4.2], and all of the narrow-channel effects can be qualitatively understood from a consideration of the depletion region cross section of this structure. This cross section is shown in Fig. 4.1, where the qualitative shape of the depletion region is shown for a gate bias near the threshold voltage. The heavy line at the silicon-silicon-dioxide interface represents the inversion channel, and the dashed lines represent the depletion region assumed by the large-geometry approximation. The increase in threshold voltage, observed as the gate width approaches the depletion depth, is produced by the excess depletion charge relative to the large-geometry approximation.
Figure 4.1 - The width cross section of the finite gate structure showing the qualitative shape of the depletion region.

The threshold voltage for a large-geometry device can be written as

\[ V_T = V_{FB} + 2\phi_F - \frac{Q_b}{C_{OX}} \]  \hspace{1cm} (4.1)

where \( Q_b \) is the bulk depletion charge per unit area and \( C_{OX} \) is the oxide capacitance per unit area. For the structure of Fig. 4.1, the flat-band voltage can be treated as a constant, independent of any geometric considerations. The \( Q_b/C_{OX} \) term above assumes a uniform depletion depth, with perfectly vertical field lines from the gate terminating in the depletion region. Fringing fields and the excess bulk charge at the channel edges are neglected.
For a narrow-channel device, the $-Q_b/C_{ox}$ term can be replaced by a similar expression which is the quotient of the total depletion charge and the total oxide capacitance per unit length for the cross section of Fig. 4.1. As an example, the excess depletion charge at the edges of the channel can be calculated using the cylindrical approximation shown in Fig. 4.2. This produces the following expression for the narrow-channel threshold voltage

$$V_T = V_{FB} + 2\phi_F + \frac{qN_A[w_xd + \frac{\pi x_d^2}{2}]}{WC_{ox}}$$

(4.2)

---

Figure 4.2 - The cylindrical approximation to the depletion region of the finite gate structure.
(If the depletion depth is much larger than the oxide thickness, the fringing capacitance can be neglected, and the total capacitance per unit length can be approximated by $WC_{ox}$.) Therefore, the added threshold voltage, $\Delta V_T$, due to the excess depletion charge is given by

$$\Delta V_T = \frac{qN_A x_d}{C_{ox}} \frac{\pi x_d}{2W} \quad (4.3)$$

where $x_d$ is the depletion depth which will vary with the bulk bias. This result will hold for long channel devices with zero drain-to-source bias, where there is no variation in the depletion charge along the channel. To include short-channel effects simultaneously with narrow-channel effects, or to include the effect of drain-to-source bias, a similar calculation can be performed where the total depletion charge and the total gate capacitance for the entire device are used. This method for obtaining the lumped threshold voltage has been used by several authors for a variety of assumed physical device structures [4.2,4.4-4.7].

Looking back at Fig. 4.1, however, it should be clear that the total depletion charge (or the lumped threshold voltage) is not the only issue for narrow-channel devices. The field lines from the gate spread in the silicon to support the excess depletion areas at the edge of the device. As a result, the surface potential is greatest at the center of the channel, where by symmetry the field is
purely vertical, and decreases towards the edges of the gate. This means that the threshold voltage is a function of position across the width of the device. As the gate voltage is increased from below threshold, the center of the transistor will invert first, and the channel edges will move outward as the gate voltage continues to increase. (Fig. 4.1 shows a gate bias for which most, but not all, of the area under the gate has been inverted.) For very large gate biases, the channel width will actually become larger than the physical gate width, with the surface inversion beyond the gate edge being supported by fringing fields. The calculation of the effective channel width versus gate voltage is a more complicated problem than the lumped threshold voltage calculation above, but this calculation nonetheless has been attempted by a few authors [4.8,4.9,4.11]. A new solution to this problem is presented below for a tapered oxide structure.

A slightly more realistic physical structure than that of Fig. 4.1 is shown in Fig. 4.3, where an abrupt transition from the thin gate oxide to the thick field oxide is assumed [4.5,4.7,4.11]. This structure has the same qualitative behavior as the uniform oxide structure discussed above. However, in this case, the excess depletion charge is a function of both the depletion depth under the gate oxide and the depletion depth under the field oxide. Unfortun-
Figure 4.3 - The width cross section of the abrupt oxide step structure.

ately, the abrupt oxide transition is hardly realistic for modern devices, particularly where the modulation of the gate width with gate bias is concerned.

The threshold voltage for a narrow-channel device assuming a linear oxide taper has been calculated by Merckel [4.4], and also by Akers et al [4.6] who included the effects of the lateral diffusion of the field implant. These models are again based on a lumped approximation to the depletion charge for a single bias point. The varying effective channel width of a narrow-channel device versus gate voltage has been calculated by Kotecha and Beilstein [4.8] who assumed a hyperbolic oxide taper rather than a
linear taper. This analysis shows that the effective channel width is different for the drain current than for the gate capacitance. A similar analysis to that of Kotecha and Beilstein is described below, using slightly different assumptions; where Kotecha and Beilstein assumed a hyperbolic oxide transition with vertical fields, a linear oxide taper with cylindrical fields has been assumed here. Although the results are similar, the analysis presented here provides a somewhat simpler solution.

The linear oxide taper structure is shown in Fig. 4.4. It is assumed that the field oxide is symmetric around the center of the gate oxide, as would approximately be the case for a non-recessed field oxidation where the ratio of the silicon consumed to the final oxide thickness is roughly one half. With the assumptions outlined below, the results depend only on the total angle between the gate and silicon surfaces and not on the individual angles of each surface relative to the plane of the silicon substrate. Therefore, with these assumptions, a fully recessed field oxide would produce the same results. The heavy line at the silicon surface again represents the surface inversion layer, which extends a distance $\Delta W$ past the corner of the oxide structure. The distance along the silicon surface is represented by the variable $\Delta x$, which is zero at the corner.
Figure 4.4 - The width cross section of the linear oxide taper structure.

The first step in calculating the width modulation is to determine the added channel width, $\Delta W$, versus the gate voltage. This is done by considering the requirements for surface inversion along the oxide taper. Given $\Delta W(V_G)$, the total inversion charge in the extended channel region can be calculated, and from this the effective channel widths for current and capacitance can be determined. The assumptions that will be used for this calculation are summarized below.

i) The electric field is assumed to be perfectly vertical under the flat part of the gate and cylindrical in the taper region. Therefore, the electric field is always perpendicular to the silicon surface.
ii) The surface potential is assumed to be constant everywhere along the inversion layer, all the way out to $AW$.

iii) The threshold voltage versus position in the taper region is determined by the same critical field at the silicon surface as that required to support inversion under the flat gate region. This means that the depletion charge associated with each incremental increase in the channel width is assumed to be independent of position. This neglects the sharing of depletion charge which will occur at the corner.

iv) The excess depletion charge at the edge of the channel is neglected. This is equivalent to assuming that the excess depletion charge can be supported by the residual electric field beyond the edge of the inversion layer. As a result, the lumped effect of the excess depletion on the threshold voltage is ignored, and the large-geometry threshold voltage is used.

v) Only the cross section of a long channel device with zero drain-to-source bias is considered. The effect of drain-to-source bias will be discussed later. The gate voltage, $V_G$, is specified relative to the source, which is assumed to be at zero potential.

The flat-band voltage for a large-geometry device is given by

$$V_{FB} = \phi_{MS} - \frac{Q_{OX}}{C_{OX}}$$

(4.4)
where $\phi_{MS}$ is the gate-semiconductor work function difference and $Q_{OX}$ is the effective interface charge density per unit area. If this is incorporated into (4.1), the threshold voltage for a large-geometry device can be rewritten as

$$V_T = (\phi_{MS} + 2\phi_F) - \frac{1}{C_{OX}} (Q_{OX} + Q_b)$$

(4.5)

The term $-(Q_{OX} + Q_b)/C_{OX}$ is the potential across the oxide at threshold, which is required to generate the critical field at the surface that will support the oxide and bulk charges. The critical oxide field can be written as

$$E_c = -\frac{(Q_{OX} + Q_b)}{\varepsilon_{OX}} = \frac{V_T - \Phi}{t_{OX}}$$

(4.6)

where $\Phi = \phi_{MS} + 2\phi_F$, and $t_{OX}$ is the oxide thickness under the flat part of the gate. The critical field will be used below, instead of the threshold voltage, as the criterion for inversion in the taper region. This is convenient because the oxide capacitance per unit area, used in the expression for $V_T$, is not constant along the taper.

The magnitude of the cylindrical field in the taper region is given by the potential across the oxide, adjusted by $\Phi$, divided by the cylindrical path length of the field lines. The pathlength is obtained by multiplying the total angle ($2\theta$), in radians, by the radius from the center of the cylindrical arc to a particular point along the oxide taper.
Therefore,

\[ E(\Delta x) = \frac{V_G - \Phi}{2 \theta (\Delta x + \frac{t_{ox}}{2 \sin \theta})} \]  

(4.7)

This expression for the electric field produces a small discontinuity at \( \Delta x = 0 \) where the cylindrical pathlength is \( \theta / \sin \theta \) times larger than \( t_{ox} \). Since this discontinuity would also introduce anomalies in \( \Delta W \), the electric field has instead been approximated by

\[ E(\Delta x) = \frac{V_G - \Phi}{2 \theta \Delta x + t_{ox}} \]  

(4.8)

Setting \( E(\Delta x) \) equal to \( E_C \) produces

\[ \Delta W = \frac{t_{ox}}{2 \theta} \left[ \frac{V_G - \Phi}{V_T - \Phi} - 1 \right] \]  

(4.9)

Before proceeding with a calculation of the effective channel widths for this case, a similar calculation will be performed to find \( \Delta W \) assuming only lateral diffusion of the field implant with no lateral oxidation. The effect of doping encroachment has been discussed by several authors [4.6, 4.9, 4.10]; however, no author has specifically addressed the question of which physical mechanism has the strongest effect on threshold voltage or channel width modulation. This issue is addressed below.
Consider the physical structure of Fig. 4.5. A uniform gate oxide is assumed and the dotted lines bound the region of constant substrate doping. Outside of this region, it is assumed that the effective substrate doping increases as

\[ N = N_A \left(1 + \frac{\Delta x}{\delta}\right) \]  \hspace{1cm} (4.10)

where \( \delta \) is the characteristic distance for the doping encroachment. This linear approximation to the lateral doping encroachment is not unreasonable considering the very long diffusion time for the field implant, which makes the lateral doping profile fairly gentle. For this struc-

![Diagram](Image)

**Figure 4.5** - The width cross section of the uniform gate structure with doping encroachment.
ture, the threshold voltage as a function of position can be described by

\[ V_T(\Delta x) = (\phi_{MS} - \frac{Q_{ox}}{C_{ox}} + 2\phi_F) - \frac{Q_b(\Delta x)}{C_{ox}} \]  
(4.11)

This can be rewritten as

\[ V_T(\Delta x) = V_A + V_B \left(1 + \frac{\Delta x}{\delta}\right)^\frac{1}{2} \]  
(4.12)

where

\[ V_A = (\phi_{MS} - \frac{Q_{ox}}{C_{ox}} + 2\phi_F) \]  
(4.13a)

and

\[ V_B = \frac{(2\varepsilon_s q N_A 2\phi_F)^\frac{1}{2}}{C_{ox}} \]  
(4.13b)

At the channel edge, the gate voltage will just equal \( V_T(\Delta x) \). Setting \( V_G \) equal to \( V_T(\Delta x) \), and solving for \( \Delta W(V_G) \), produces

\[ \Delta W = \delta \left[ \left( \frac{V_G - V_A}{V_B} \right)^2 - 1 \right] \]  
(4.14)

So the added channel width for the doping encroachment case increases as \( V_G^2 \) rather than as \( V_G \) for the oxide encroachment case. Therefore, the channel width modulation is likely to be limited by the increasing oxide thickness in the taper region rather than by the laterally increasing substrate doping.
This can be demonstrated by a numerical example. For the CMOS process used here, the field oxide thickness is 1.3 μm, and the field threshold is approximately 30 volts. Assuming an interface state density \( \frac{Q_{OX}}{q} \) of \( 1 \times 10^{11} \), the effective substrate doping in the field region must be approximately \( 4 \times 10^{16} \), which is only five times the native substrate concentration of \( 8 \times 10^{15} \). This increases the \( -\frac{Q_b}{C_{OX}} \) term in the threshold voltage by only a little more than one volt. So, without the increase in the oxide thickness, a one volt increase in \( V_G \) would extend the channel width all the way under the field region. Clearly then, the channel width modulation must be limited by the oxide taper and not the lateral doping profile. (This analysis assumes that the substrate doping is uniform into the silicon, i.e., that there is no threshold voltage tailoring implant. This assumption has been made throughout. If this is not the case, a more complicated expression for the threshold voltage is required.)

Returning to the oxide taper case of Fig. 4.4 and (4.9), the key to calculating the effective channel widths is the calculation of the total channel charge in the extended width region. The channel charge, as a function of \( \Delta x \), can be determined by the electric field at the silicon surface in excess of the critical field as follows

\[
\Delta Q_n(\Delta x) = -\varepsilon_{OX} [E(\Delta x) - E_c]
\]  

(4.15)
The total channel charge in the extended width region is therefore given by

\[
\Delta Q_n = \int_0^{\Delta W} \Delta Q_n(\Delta x) \, d\Delta x
\]

\[
= - \varepsilon_{ox} \int_0^{\Delta W} \left[ \frac{V_G - \Phi}{2 \theta \Delta x + t_{ox}} - \frac{V_T - \Phi}{t_{ox}} \right] \, d\Delta x \tag{4.16}
\]

The evaluation of (4.16) produces

\[
\Delta Q_n = - \varepsilon_{ox} \frac{2}{\theta} \left[ (V_G - \Phi) \ln \left( \frac{V_G - \Phi}{V_T - \Phi} \right) - (V_G - V_T) \right] \tag{4.17}
\]

The effective current width, \( W_{Id} \), for a cross section of the transistor, is proportional to the total channel charge per unit length of the cross section. The total channel charge is given by

\[
Q_n = \bar{Q}_n + 2\Delta Q_n = - \bar{W} \varepsilon_{ox} (V_G - V_T) + 2\Delta Q_n \tag{4.18}
\]

where \( Q_n \) is the channel charge in the flat gate region of width \( \bar{W} \), which is treated as for a large-geometry device. Combining (4.17) and (4.18) and factoring out \( \bar{Q}_n \) produces

\[
Q_n = \bar{Q}_n \left[ 1 + \frac{t_{ox}}{\bar{W} \theta} \left( \frac{V_G - \Phi}{V_G - V_T} \ln \frac{V_G - \Phi}{V_T - \Phi} - 1 \right) \right] \tag{4.19}
\]

From (4.19), the effective width for the transport current, \( W_{Id} \), can be identified as

- 171 -
\[ W_{Id} = \bar{W} \left[ 1 + \frac{t_{ox}}{W} \frac{1}{\theta} \left[ \frac{V_G}{V_G - V_T} \ln \frac{V_G}{V_T - \Phi} - 1 \right] \right] \]  \hspace{1cm} (4.20)

The gate-to-channel and bulk-to-channel capacitances are calculated from the channel charge, assuming that these capacitances are reciprocal with the channel-to-gate and channel-to-bulk capacitances respectively. This is done because of the difficulty of calculating the gate and bulk charges as required to calculate these capacitances directly. While the channel charge is non-zero only out to \( \Delta x = \Delta W \), the gate and bulk charges are appreciable all along the gate and silicon surfaces, well past \( \Delta W \). In addition, the silicon surface potential must be estimated along the oxide taper in order to calculate the gate and bulk charges. Therefore, the added gate-to-channel capacitance per unit length, \( \Delta C_g \), in the channel extension is estimated by

\[ \Delta C_g = - \frac{dQ_N}{dV_G} = \frac{\varepsilon_{ox}}{2\theta} \ln \frac{V_G}{V_T - \Phi} \]  \hspace{1cm} (4.21)

and the total gate capacitance is given by

\[ C_g = \bar{C}_g + 2\Delta C_g = \bar{W} \frac{\varepsilon_{ox}}{t_{ox}} + 2\Delta C_g \]  \hspace{1cm} (4.22)

The effective width for the gate-to-channel capacitance, \( W_{Cg} \), can be found by combining (4.21) and (4.22), and factoring out \( C_g \) as was done for \( W_{Id} \) above. This produces
\[ W_{Cg} = \bar{W} \left[ 1 + \frac{t_{ox}}{W} \frac{1}{\theta} \ln \left( \frac{V_G - \Phi}{V_T - \Phi} \right) \right] \quad (4.23) \]

The added bulk capacitance per unit length, \( \Delta C_b \), is similarly calculated by

\[ \Delta C_b = -\frac{d\Delta Q_N}{dV_B} = \frac{\varepsilon_{ox}}{2\theta} \left( \frac{V_G - \Phi}{V_T - \Phi} - 1 \right) \frac{\partial V_T}{\partial V_B} \quad (4.24) \]

From (4.5) we find

\[ \frac{\partial V_T}{\partial V_B} = -\frac{1}{C_{ox}} \frac{\partial Q_B}{\partial V_B} \quad (4.25) \]

Therefore, \( \Delta C_b \) can be written as

\[ \Delta C_b = \frac{t_{ox}}{2\theta} \left( \frac{V_G - \Phi}{V_T - \Phi} - 1 \right) \frac{\partial Q_B}{\partial V_B} \quad (4.26) \]

and total bulk capacitance is given by

\[ C_b = \bar{C}_b + 2\Delta C_b = \bar{W} \frac{\partial Q_B}{\partial V_B} + 2\Delta C_b \quad (4.27) \]

Following the same procedure as above, the effective width for the bulk capacitance, \( \bar{W}_{Cb} \), is given by

\[ \bar{W}_{Cb} = \bar{W} \left[ 1 + \frac{t_{ox}}{W} \frac{1}{\theta} \left( \frac{V_G - \Phi}{V_T - \Phi} - 1 \right) \right] \quad (4.28) \]
Notice that while the effective width for gate capacitance increases less than linearly with \( V_G \), the effective width for bulk capacitance increases linearly.

Although the lateral diffusion of the field implant was shown to be of minor importance in calculating \( \Delta W \), doping encroachment does have a significant impact on \( \Delta C_b \). In particular, as \( \Delta W \) increases, \( \Delta C_b \) increases disproportionately faster because of the lateral variation in the effective substrate doping. This effect can be calculated by noticing from (4.26) that

\[
\Delta C_b = \Delta W \frac{\partial Q_b}{\partial V_B} \tag{4.29}
\]

which can be generalized to

\[
\Delta C_b = \int_0^{\Delta W} \frac{\partial Q_b}{\partial V_B} \, d\Delta x \tag{4.30}
\]

For the doping encroachment case, (4.30) is evaluated using

\[
\frac{dQ_b(\Delta x)}{dV_B} = (1+\frac{\Delta x}{\delta})^\frac{3}{2} \frac{dQ_b}{dV_B}, \tag{4.31}
\]

which reflects the increasing bulk capacitance per unit area as the effective substrate doping increases with increasing \( \Delta x \). This produces

\[
\Delta C_b = \frac{2}{3} \delta \left[ \left( 1 + \frac{t_{ox}}{2\delta} \frac{V_G - \Phi}{V_T - \Phi} - 1 \right)^{3/2} - 1 \right] \tag{4.32}
\]
Calculating the effective width for the bulk capacitance with doping encroachment yields

$$W_{C_b} = \frac{4}{3} \frac{\delta}{W} \left[ 1 + \frac{t_{ox}}{208} \frac{V_G - \Phi}{V_T - \Phi} \left( 1 + \frac{t_{ox}}{208} \frac{V_G - \Phi}{V_T - \Phi} - 1 \right) \right]^{3/2}$$  \hspace{1cm} (4.33)

The interesting feature of (4.33) is that it predicts a faster than linear increase of the bulk capacitance with $V_G$. This was experimentally observed, as noted in the next section. Also notice that (4.33) reduces to (4.28) if $\delta$ is very large, meaning that the slope of the lateral doping profile is very small.

The effective widths calculated above only apply for a cross-section of the MOSFET with zero drain-to-source bias. (The effect of substrate bias is included in $Q_d$ and $V_T$.) In order to correctly calculate the total effective widths for the device, the cross-sectional behavior would have to be included in the channel charge expression used in the basic transport equation, and the transport current and channel charge expressions would have to be integrated from source to drain to obtain the total device current and capacitances. This procedure is necessary to completely account for the varying surface potential and depletion depth along the channel.

Instead, a simple approximation for the effective widths is introduced in section 4.1.3. This model accur-
ately predicts the narrow-channel effects observed in the experimental data. The next section reviews the experimental capacitance data, and identifies the various narrow-channel effects. In addition, the narrow-channel effects on the drain current are noted and are related to the narrow-channel capacitance phenomena.
4.1.2 Observations on the experimental data

Narrow-channel capacitance effects can be observed by comparing the experimental data from the 4 μm by 20 μm and 20 μm by 20 μ (W/L) devices, Figs. A.6 and A.8 respectively. Although the 20 μm by 20 μm device exhibits subtle small-geometry effects relative to the 100 μm by 100 μm device, it is necessary to compare two devices with the same channel length in order to isolate the narrow-channel phenomena. The differences between the 4 μm wide device data and the 20 μm wide device data, for each of the device capacitances, are summarized below.

\( C_{gs} \)

i) There is an upward tilt to the \( C_{gs} \) curves above \( V_T \) for the narrow device. (This can be most easily seen for the \( V_{DS} = 0 \) curve.)

ii) There is no appreciable effect on the extrinsic capacitances, i.e., there is no voltage dependence of the capacitance curves below \( V_T \).

\( C_{gd} \)

i) There is an upward tilt to the \( C_{gd} \) curves above \( V_T \) for the narrow device.

ii) There is no appreciable effect on the extrinsic capacitances (below \( V_T \)).
\( C_{gb} \)
i) The transition of \( C_{gb} \) from inversion to depletion to accumulation (\( V_{GS} \) decreasing below \( V_T \)) is less abrupt for the narrow device than for the wide device. (This can be observed from both the intrinsic and extrinsic plots.)

ii) There is a voltage-dependent contribution to the extrinsic part of \( C_{gb} \) which decreases as \( V_{GS} \) increases. (This can be seen from the \( V_{DS} = 0 \) curve of the intrinsic plot.) This extrinsic component has a slight dependence on \( V_{DS} \), indicated by the splitting of the \( C_{gb} \) curves at \( V_{GS} - V_T = 6 \) \( V \).

\( C_{gg} \)
i) The transition of \( C_{gg} \) from depletion to accumulation is less abrupt for the narrow device.

ii) The depth of the notch in the curves at depletion is less for the narrow device. (This may be a side effect of the normalization calculations.)

iii) The shape of the intrinsic \( C_{gg} \) curves for the narrow and wide devices are roughly the same. This indicates that the added capacitance of \( C_{gs} \) plus \( C_{gd} \), as \( V_{GS} \) increases, is approximately balanced by decrease in the extrinsic capacitance of \( C_{gb} \).
\[ C_{bs} \]

i) There is an upward tilt to the \( C_{bs} \) curves above \( V_T \) for the narrow device.

ii) The intrinsic \( C_{bs} \) capacitances are larger for the narrow-channel device.

iii) There is no appreciable effect on the extrinsic capacitances (below \( V_T \)).

\[ C_{bd} \]

i) There is an upward tilt to the \( C_{bd} \) curves above \( V_T \) for the narrow device.

ii) The intrinsic \( C_{bd} \) capacitances are larger for the narrow-channel device.

iii) There is no appreciable effect on the extrinsic capacitances (below \( V_T \)).

\[ C_{bg} \]

i) The transition of \( C_{bg} \) from inversion to depletion to accumulation (\( V_{GS} \) decreasing below \( V_T \)) is less abrupt for the narrow device than for the wide device.

ii) There is a voltage-dependent contribution to the extrinsic part of \( C_{bg} \) which decreases as \( V_{GS} \) increases. (This can be seen from the \( V_{DS} = 0 \) curve of the intrinsic plot.) This extrinsic component has a slight dependence on \( V_{DS} \), indicated by the splitting of the \( C_{bg} \) curves at \( V_{GS} - V_T = 6 \) v.
i) The transition of $C_{bb}$ from depletion to accumulation is less abrupt for the narrow device.

ii) The depth of the notch in the curves at depletion is less for the narrow device. (This may be a side effect of the normalization calculations.)

iii) The intrinsic $C_{bb}$ curves for the narrow device increase with increasing $V_{GS}$. This indicates that the added capacitance of $C_{BS}$ plus $C_{bd}$, as $V_{GS}$ increases, is larger than the decrease in the extrinsic part of $C_{gb}$.

In addition to these specific observations, there are several characteristics which are common to all of the device capacitances.

i) The turn-on transitions near $V_T$ are less abrupt for the narrow-channel device than for the wide device.

ii) The threshold voltage shifts of the $V_{BS} = -2v$ extrinsic curves relative to the $V_{BS} = 0v$ extrinsic curves are larger for the narrow-channel device.

iii) The drain-to-source saturation voltage ($V_{DSSat}$) is smaller for the narrow-channel devices. This can be seen by a shift to the right of the upward transitions in the $C_{gd}$ and $C_{gg}$ curves.
The most significant narrow-channel capacitance effect is the increase in the intrinsic parts of $C_{gs}$, $C_{gd}$, $C_{bs}$, and $C_{bd}$, and the decrease in the extrinsic parts of $C_{gb}$ and $C_{bg}$ as $V_{GS}$ increases. This is indicative of the increasing channel width for increasing gate-to-source voltage. (The decrease in the extrinsic parts of $C_{gb}$ and $C_{bg}$ occurs because the increasing channel width reduces the capacitive coupling from gate to bulk around the edges of the device.) The other narrow-channel effects (the increased shift of threshold voltage with source-to-bulk bias, the decreased drain-to-source saturation voltages, and the sluggish depletion-accumulation transition) are all consistent with a higher effective substrate doping or body effect factor ($\gamma$). The larger body effect factor is related to the excess depletion charge of the narrow-channel device which gives rise to the positive shift of threshold voltage. These same effects can be observed in the drain current data, and in the electrical parameters typically extracted from this data.

These electrical parameters are shown in Table 4.1, where data for 4 \( \mu \text{m} \) wide devices are compared to data for 20 \( \mu \text{m} \) wide devices. (These data do not come directly from the drain current curves of Figs. A.6 and A.8; rather, they are the average values of measurements from several devices.) An increase in the electrical device width for increasing gate-to-source voltage is indicated by the
Table 4.1

Electrical Parameters of the
20 μm and 4 μm Channel Width Devices

<table>
<thead>
<tr>
<th>Electrical Parameter</th>
<th>20 μm Width</th>
<th>4 μm Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effective Channel Length (V_{GS} ≈ 1V)</td>
<td>17.1 μm</td>
<td>1.1 μm</td>
</tr>
<tr>
<td>Threshold Voltage</td>
<td>0.72 V</td>
<td>1.20 V</td>
</tr>
<tr>
<td>Body Effect Factor (γ)</td>
<td>1.07 V^2</td>
<td>1.47 V^2</td>
</tr>
<tr>
<td>Mobility Degradation Factor (θ)</td>
<td>0.04 V^{-1}</td>
<td>-0.01 V^{-1}</td>
</tr>
</tbody>
</table>

reduced (actually slightly negative) mobility degradation factor, θ, for narrow-channel devices. This parameter is extracted from a plot of drain current versus V_{GS} for a small fixed drain-to-source voltage. (Fig. A.8a is an example of this type of plot.) Ideally, for a large-geometry device with no mobility degradation, this plot of the drain current should be a straight line with a slope that is proportional to the electron mobility. In practice, the slope of the curve decreases with increasing V_{GS}, reflecting the reduced electron mobility for large vertical fields at the silicon surface. The mobility degradation
parameter is extracted from the changing slope of the $V_{SB} = 0V$ curve assuming a mobility model of the form

$$\mu = \mu_0 \frac{1}{1 + \theta(V_{GS} - V_T)} .$$

(This expression approximates the model of White et al [4.12] for the case of $V_{SB} = 0V$.) For the narrow-channel device, the tendency of the drain current to increase less than linearly with $V_{GS}$ due to mobility degradation is more than compensated for by an increasing effective channel width. From the difference in the extracted mobility degradation factors, the effective channel width of the narrow-channel device appears to be increasing by approximately 5 \% for every one volt increase in $V_{GS}$.

The apparent increase in the effective substrate doping or body effect in the narrow-channel capacitance data is consistent with the larger threshold voltage and body effect factor ($\gamma$) for narrow-channel devices obtained from the drain current data. (The threshold voltage is determined by an extrapolation of the linear drain current characteristic, at the point of maximum slope, back to the x-axis, and gamma ($\gamma$) is extracted from the variation of $V_T$ with source-to-bulk bias.) So, the same phenomena seem to be present, at least qualitatively, in both the capacitance and static drain current characteristics. Moreover, these phenomena are consistent with the theoretical analysis of the preced-
ing section. In the next section, simple models for the
gate and bulk capacitances are described which are based on
the theoretical analysis above, and which accurately predict
the narrow-channel capacitance data.
4.1.3 Modeling narrow-channel capacitances

There are two steps which must be taken to accurately model narrow-channel capacitance effects, and the first step is choosing an appropriate narrow-channel threshold voltage model. The most complete narrow-channel threshold voltage model appears to be that of Akers et al [4.6], which includes the effect of a linear oxide taper and doping encroachment. From the substrate voltage dependence of the threshold voltage model, an effective body effect factor can be calculated. (The body effect factor of a narrow-channel device is larger than that of a large-geometry device due to the influence of the excess bulk charge at the channel edges.) Using the narrow-channel threshold voltage and body effect factor ($\gamma$) in a large-geometry capacitance model (such as those of Meyer [4.13] or Ward [4.14]) will accurately model many of the narrow-channel capacitance effects: the larger substrate capacitances, the reduced drain-to-source saturation voltage, and the less abrupt transitions from depletion to accumulation and from depletion to inversion. The only difficulty in this procedure is the calculation of a representative body effect factor from the narrow-channel threshold voltage expression.

For a large-geometry device, the threshold voltage (at the source) can be written as
\[ V_T = V_{FB} + 2\phi_F + \gamma(V_{SB} + 2\phi_F)^\lambda \]  \hspace{1cm} (4.35)

\[
\gamma(V_{SB} + 2\phi_F)^\lambda = -\frac{Q_b}{C_{ox}} = \frac{qN_A x_d}{C_{ox}} \hspace{1cm} (4.36)
\]

and \(x_d\) is the junction depth for a voltage drop of \(V_{SB} + 2\phi_F\) across the depletion region. For the simple finite gate structure of Fig. 4.1, the narrow-channel threshold voltage was calculated above and is given by

\[
V_T = V_{FB} + 2\phi_F + \frac{qN_A [Wx_d + \frac{\pi x_d^2}{2}]}{WC_{ox}} \hspace{1cm} (4.37)
\]

Because of the dependence of the narrow-channel threshold voltage on \(x_d^2\), a single bias-independent body effect factor can not be extracted from (4.37). One alternative to using an approximate bias-independent value for \(\gamma\) is to define a bias-dependent body effect factor, which would be chosen to fit (4.37) at the median potential along the channel, i.e., \(V_{SB} + V_{DS}/2\). A bias-dependent body effect factor could be used in either a charge-oriented model or a capacitance model, and would allow the accurate modeling of the device capacitances which are sensitive to body effect, such as \(C_{bs}, C_{bd}, C_{gb}\), and \(C_{bg}\).

The second step in modeling narrow-channel capacitances is to incorporate the phenomenon of increasing effective channel width for increasing gate voltage. As discussed in
section 4.1.1, the effective width for the transport current differs from the effective width for the gate capacitances which, in turn, differs from the effective width for the bulk capacitances. Before the effective widths calculated in section 4.1.1 can be used for capacitance modeling, there are two adjustments which must be made.

The first adjustment concerns the use of the large-geometry threshold voltage expression in the narrow-channel analysis. It was assumed above that the area under the flat part of the gate is fully inverted when the gate is at the large-geometry threshold voltage. In addition, the large-geometry threshold voltage was used in the calculation of the critical field for inversion along the oxide taper. This assumes that the same amount of bulk charge per unit area must be supported by the surface field for the narrow-channel case as for the large-geometry case. With respect to the first assumption, the correct threshold voltage for a narrow-channel device must be used where the threshold voltage appears as a reference for the gate voltage relative to inversion. For the second assumption, the large-geometry threshold voltage can be used for the critical field since fringing fields were ignored. Making these adjustments, the channel width along the oxide taper, \( \Delta W \), can be written as

\[
\Delta W = \frac{t_{ox}}{2\theta} \left[ \frac{V_G - V_T}{V_b} \right]
\]  

(4.38)
where

\[ V_b = -\frac{(Q_{ox} + Q_D)}{C_{ox}} \]  

(4.39)

and \( Q_D \) is the bulk charge per unit area for a large-geometry device. The key feature of (4.38) is that it uses only the difference between the gate and threshold voltages, so that any value for the threshold voltage can be used independent of the constant associated with the bulk (and oxide) charges. Similar modifications of the effective width terms will be made below.

The second adjustment corrects for the assumption of zero drain-to-source bias in section 4.1.1. Given a drain-to-source bias, the value of \( \Delta W \) will vary along the channel as the surface quasi-Fermi potential varies from \( V_S \) to \( V_D \). Strictly speaking, the effect of the varying channel width should be integrated from source to drain in order to calculate the total effective channel width. Rather than carrying out this integral, a simple approximation is used here. Specifically, a weighted average of the \( \Delta W \)'s evaluated at the source and drain is used, as shown below:

\[ \Delta W_{\text{eff}} = \frac{2}{3} \Delta W(V=V_S) + \frac{1}{3} \Delta W(V=V_D) \]  

(4.40)

Since the quasi-Fermi potential varies more slowly near the source and more quickly near the drain, the weighting factor is chosen to emphasize the width at the source over the
width at the drain. The choice of 2/3 for the weighting factor for the source term is suggested by the 2/3 that appears in the expression for the gate-to-source capacitance in saturation.

Including these adjustments in (4.23), the intrinsic gate-to-channel capacitances, $C_{gs}$ and $C_{gd}$, can be approximated by

$$C_{gs} = \bar{C}_{gs} \left[ 1 + \frac{2}{3} \frac{1}{W} \Delta W_g(V=V_S) + \frac{1}{3} \frac{1}{W} \Delta W_g(V=V_D) \right]$$

and

$$C_{gd} = \bar{C}_{gd} \left[ 1 + \frac{2}{3} \frac{1}{W} \Delta W_g(V=V_S) + \frac{1}{3} \frac{1}{W} \Delta W_g(V=V_D) \right]$$

(4.41)

where $\bar{C}_{gs}$ and $\bar{C}_{gd}$ are the large-geometry capacitances calculated using $W = \bar{W}$ and using the effective narrow-channel body effect factor discussed above, and where

$$\Delta W_g(V=V_S) = \frac{t_{ox}}{\theta} \ln \left[ \frac{V_{GS}-V_T}{V_b(V_{SB})} + 1 \right]$$

and

$$\Delta W_g(V=V_D) = \frac{t_{ox}}{\theta} \ln \left[ \frac{V_{GD}-V_T}{V_b(V_{DB})} + 1 \right].$$

(4.42)

In saturation, $V_{Dsat}$ is used for $V_D$ in $\Delta W_g(V=V_D)$, which makes $\Delta W_g=0$.

Similarly, from (4.33), the intrinsic bulk-to-channel capacitances, $C_{bs}$ and $C_{bd}$, can be approximated by
\[
C_{bs} = \bar{C}_{bs} \left[ 1 + \frac{2}{3} \frac{1}{w} \Delta W_b(V=V_S) + \frac{1}{3} \frac{1}{w} \Delta W_b(V=V_D) \right]
\]
and
\[
C_{bd} = \bar{C}_{bd} \left[ 1 + \frac{2}{3} \frac{1}{w} \Delta W_b(V=V_S) + \frac{1}{3} \frac{1}{w} \Delta W_b(V=V_D) \right]
\]
where
\[
\Delta W_b(V=V_S) = \frac{4}{3} \delta \left[ (1 + \frac{t_{ox}}{20 \delta} \frac{V_{GS}-V_T}{V_b(V_{SB})})^{3/2} - 1 \right]
\]
and
\[
\Delta W_b(V=V_D) = \frac{4}{3} \delta \left[ (1 + \frac{t_{ox}}{20 \delta} \frac{V_{GD}-V_T}{V_b(V_{DB})})^{3/2} - 1 \right].
\]

Figures 4.6-4.9 demonstrate the accuracy of the bias-dependent effective-width models for narrow-channel devices. Rather than calculating the large-geometry capacitances from the Meyer [4.13] or Ward [4.14] models for use in (4.41) and (4.43), the experimental 20 μm by 20 μm capacitance data were used. This ensures that no channel length effects interfere with the comparison of the experimental and calculated narrow-channel data. Unfortunately, the narrow-channel data calculated from the 20 μm by 20 μm device data do not reflect the increased body effect for the narrow-channel device. Therefore, the agreement is best for the \( V_{DS} = 0 \) v curves.

Figures 4.6 and 4.7 show the experimental and calculated data for the narrow-channel gate-to-channel capacitances (\( C_{gs} \) and \( C_{gd} \) respectively.) The solid lines are the experimental data, and the dotted lines are the capacitances
Figure 4.6 - Calculated and experimental intrinsic $C_{gs}$ of a 4 $\mu$m by 20 $\mu$m (W/L) device, normalized by $C_{oi}$.

Figure 4.8 - Calculated and experimental intrinsic $C_{ds}$ of a 4 $\mu$m by 20 $\mu$m (W/L) device, normalized by $C_{oi}$.
calculated by using the 20 μm by 20 μm data in (4.31). Excellent agreement was achieved, particularly for $V_{DS} = 0\text{v}$, using $\theta = 0.58$ radians (33 degrees) and $W = 1.1$ μm. (The gate oxide thickness is 700 Å.) This value for the taper angle is in good agreement with studies of two-dimensional oxidation [4.15]. With these values, the effective channel width for the gate capacitances is calculated to be 1.26 W for $V_{GS} - V_T = 6\text{v}$ and $V_{DS} = 0\text{v}$. Much of the disagreement for larger drain-to-source voltages is attributable to the mismatch of the drain-to-source saturation voltages between the 20 μm and the 4 μm channel width devices.

Figures 4.8 and 4.9 show the experimental and calculated data for the narrow-channel bulk-to-channel capacitances ($C_{bs}$ and $C_{bd}$ respectively.) Excellent agreement was again obtained, this time using $\delta = 0.22$ μm and $\theta = 1.00$ radians (58 degrees). The value of $\delta$ expected from geometrical considerations is 0.23, assuming a taper angle of 45 degrees and a variation of the effective substrate doping from $N_A$ to $5N_A$ along the oxide taper. It is not clear if there is a fundamental reason why there should be a discrepancy in the effective taper angles for the gate and bulk capacitances. It was noticed, however, that the curve fitting was not very sensitive to variations in $\theta$. Since only simple curve fitting techniques were used, an equally good fit may have been attainable with substantially different parameters. In any event, the qualitative agreement is
Figure 4.8 - Calculated and experimental intrinsic $C_{bs}$ of a 4 μm by 20 μm (W/L) device, normalized by $C_{oi}$.

Figure 4.9 - Calculated and experimental intrinsic $C_{bd}$ of a 4 μm by 20 μm (W/L) device, normalized by $C_{oi}$. 
very good. It should be noted that the absolute scale factor had to be adjusted to account for the difference in body effect factors, since the substrate capacitances are directly proportional to γ. The scale factor chosen for the best fit to the experimental data was very near the value predicted by the ratio of γ's from Table 4.1.

Finally, a bias dependence in the extrinsic parts of \( C_{gb} \) and \( C_{bg} \) was observed from the experimental data. This can be modeled by assuming that the capacitance gained between gate and channel is lost from gate to bulk. Therefore, \( C_{gbe} \) and \( C_{bge} \) can be approximated by

\[
C_{gbe} = \overline{C}_{gbe} - \overline{C}_{ox} \left[ \frac{1}{3} \frac{1}{W} \Delta W_g(V=V_S) + \frac{1}{3} \frac{1}{W} \Delta W_g(V=V_D) \right]
\]

and

\[
C_{bge} = \overline{C}_{bge} - \overline{C}_{ox} \left[ \frac{1}{3} \frac{1}{W} \Delta W_g(V=V_S) + \frac{1}{3} \frac{1}{W} \Delta W_g(V=V_D) \right]
\]

(4.45)

where \( \overline{C}_{ox} = WLC_{ox} \), and where \( \overline{C}_{gbe} \) and \( \overline{C}_{bge} \) include the total capacitance of the oxide taper. In other words, as the gate voltage becomes very large, the terms subtracted above should cancel the taper capacitance which is included at low gate voltages. Figures 4.10 and 4.11 compare the experimental gate-to-bulk and bulk-to-gate capacitances to the calculated capacitances obtained from the 20 μm by 20 μm data as above. The agreement is again excellent, even for the splitting of the curves for large gate voltages due to the drain dependence of the extrinsic capacitances. The
Figure 4.10 - Calculated and experimental intrinsic $C_{gb}$ of a 4 µm by 20 µm (W/L) device, normalized by $C_{oi}$.

Figure 4.11 - Calculated and experimental intrinsic $C_{bg}$ of a 4 µm by 20 µm (W/L) device, normalized by $C_{oi}$. 

- 195 -
disagreement below $V_T$ occurs because the scale factor has again been adjusted to match the intrinsic capacitances (above $V_T$) which are proportional to $\gamma$. This throws off the matching of the accumulation capacitance below $V_T$.

In conclusion, simple models for the bias dependence of the effective channel width for the capacitances of narrow-channel devices have been presented. These models show good agreement with the experimental data. With the addition of an appropriate threshold voltage and body effect model, the effective channel width models should provide an effective means of modeling narrow-channel device capacitances using an underlying large-geometry capacitance model.
4.2 Short-Channel Capacitance Effects

There are two separate phenomena which must be considered for short-channel devices. The first is the decreased threshold voltage due to depletion charge sharing between the gate and the source and between the gate and the drain near the ends of the channel. This phenomenon has been studied by many authors [4.16-4.30], and it can be understood by considering the total depletion charge for a cross section along the length of the device. This analysis is similar to that for the narrow channel threshold voltage effect. The second short-channel phenomenon is that of velocity-limited saturation [4.31-4.39]. Short-channel devices saturate with a smaller drain-to-source voltage because of the finite velocity of the channel carriers for high lateral fields. Velocity-limited saturation alters the distribution of charge along the channel and therefore has a significant impact on the intrinsic capacitances. As in the previous section, these basic phenomena will first be reviewed, and then the experimental capacitance data will be examined to identify the short-channel effects. Finally, simple models are proposed for the capacitance effects caused by charge sharing. A complete analysis of the effect of velocity saturation on the device capacitances, however, is not attempted here.
4.2.1 Theoretical basis for short-channel effects

Fig. 4.12 represents a cross section along the length of a short-channel device and shows the qualitative depletion region with some drain-to-source bias. The dotted lines represent the depletion regions which would be produced by the source and drain junctions in the absence of any gate bias. The threshold voltage is reduced relative to that of a long channel device because of this overlap of the channel and junction depletion regions; some of the bulk charge that would normally be supported by field lines from the gate will actually be supported by field lines from the source and drain.

![Image of a cross section of a short-channel device](image)

**Figure 4.12** - The length cross section of a short-channel device showing the qualitative shape of the depletion region.
A very simple calculation of this effect was made by Varshney in 1973 [4.16]. He assumed rectangular junctions of infinite depth as well as zero drain-to-source bias. Given these assumptions, the bulk charge can be approximated by the simple rectangular regions of Fig. 4.13. By symmetry, it is reasonable to assume that one half of the charge in the square overlap regions will be mirrored at the gate and one half of this charge will be mirrored at the source or drain. (This assignment of the bulk charge between the gate and the source or between the gate and the drain is referred to as charge sharing.) With this splitting of the

![Diagram](image)

Figure 4.13 - The length cross section of a short-channel device with idealized junctions.
bulk charge in mind, and recalling (4.1), the threshold voltage for a short-channel device can be written as

\[ V_T = V_{FB} + 2\varphi_F + \frac{qN_A(Lx_d - x_d^2)}{LC_{ox}}, \quad (4.46) \]

so that the threshold voltage is reduced by

\[ \Delta V_T = -\frac{qN_Ax_d}{C_{ox}} \frac{x_d}{L}. \quad (4.47) \]

A more accurate calculation of the bulk charge, which takes into account the curvature of the source and drain junctions, was made by Yau [4.17]. Subsequently, several other authors have used similar methods to calculate the effect of drain bias on the short-channel threshold voltage [4.18-4.24]. Other methods have also been used to calculate the short-channel threshold voltage. Most of these are based on the variation of the surface potential along the channel and the resulting inversion charge [4.25-4.29]. These methods generally use the inversion charge or the surface potential as a criterion for inversion, rather than using a lumped analysis of the bulk charge at threshold. Finally, at least one author has proposed an empirical threshold voltage model based on the results of two-dimensional numerical simulations [4.30]. We will not be concerned here with any particular threshold voltage model; rather we will only be concerned with the capacitance

- 200 -
effects associated with the reduced threshold voltage and the reduced body effect.

For large-geometry devices, the drain current saturates when the drain end of the channel becomes pinched off. This occurs when the drain voltage is raised to the point where the gate-to-drain voltage is not large enough to support inversion. If the drain voltage is raised beyond this point, the potential at the drain end of the channel remains fixed and the drain voltage in excess of the pinch-off voltage is dropped across a space-charge layer, frequently referred to as the drain region. In terms of the drain-to-source voltage, the pinch-off point is defined by

\[ V_{DSSat} = V_GS - V_T \] (4.48)

where \( V_T \) is evaluated at the drain. Because of the implicit dependence of \( V_T \) on \( V_{DSSat} \), the explicit solution for \( V_{DSSat} \) is somewhat complex [4.40].

This simple theory predicts that the magnitude of the lateral electric field at the drain must become infinite at pinch-off. This is required by current continuity; since the mobile channel charge is assumed to be zero at the drain, the field (and hence the velocity of the carriers) must become infinite in order to support the transport current. In reality, the velocity-field relationship for electrons (or holes) saturates towards a finite carrier
velocity for large longitudinal electric fields (>10^4 V/cm). As a result, the drain current saturates before \( V_{DSSat} \) is reached as the channel electrons reach their saturation velocity. The discrepancy between the pinch-off voltage and the actual drain-to-source saturation voltage increases for short-channel devices which have larger lateral fields.

Beyond the point in the channel where the carriers reach their saturation velocity, the surface potential will continue to increase until the field at the silicon surface becomes too small to constrain the carriers to move in a surface channel. At this point, the drain region will form in much the same way as predicted by the simple saturation theory. The carriers will spread into the bulk, and the physical width of the drain region will be determined by the two-dimensional distribution of the mobile carriers and the fixed bulk charge. The principal consideration for the modeling of the static current in saturation is the modulation of the effective channel length with drain-to-source voltage. The effective channel length is reduced by the width of the drain region, so that modulation of the width of the drain region with variations in the drain voltage produces a non-zero output conductance.

The effects of a saturating velocity-field relationship on the drain current and on the distribution of carriers in the channel was recognized by Hofstein and Warfield in 1965
[4.31]. Subsequently, other authors have focussed on the impact of the choice of a velocity-field model on the device characteristics [4.36, 4.38] and the modulation of the drain region as it affects the output conductance in saturation. The modulation of the drain region is of particular interest for capacitance modeling since the charge in the drain region is appreciable for a velocity-saturated device. The assignment of this charge to the device terminals is another two-dimensional charge sharing question which will not be addressed quantitatively here. However, the qualitative effects of velocity saturation will be discussed below.
4.2.2 Observations on the experimental data

Short-channel capacitance effects can be observed by comparing the experimental data from the 20 μm by 4 μm and 20 μm by 20 μm (W/L) devices, Figs. A.6 and A.5 respectively. Although the 20 μm by 20 μm device exhibits subtle small-geometry effects relative to the 100 μm by 100 μm device, it is necessary to compare two devices with the same channel width in order to isolate the short-channel phenomena. The differences between the 4 μm long device data and the 20 μm long device data, for each of the device capacitances, are summarized below

\[ C_{gs} \]

i) The intrinsic part of \( C_{gs} \) is larger in saturation for the short-channel device than for the long channel device.

ii) The intrinsic part of \( C_{gs} \) in saturation increases with \( V_{GS} \), producing a curved family of drain-to-source bias curves.

iii) There is an appreciable gate voltage dependence of the extrinsic part of \( C_{gs} \) (below \( V_T \)).

iv) There is an appreciable bulk voltage dependence of the extrinsic part of \( C_{gs} \) (below \( V_T \)). (This can be seen from the extrinsic plot.)
\(C_{gd}\)  
i) There is appreciable gate-to-drain capacitance for
the short-channel device for low gate voltages
before the transistor comes out of saturation.

ii) There is an appreciable gate voltage dependence of
the extrinsic part of \(C_{gd}\) (below \(V_T\)).

iii) There is an appreciable bulk voltage dependence of
the extrinsic part of \(C_{gd}\) (below \(V_T\)).

iv) There is a slight drain voltage dependence of the
extrinsic part of \(C_{gd}\) (below \(V_T\)).

\(C_{gb}\)

i) The intrinsic \(C_{gb}\) capacitance (above \(V_T\)) is reduced
for the short-channel device.

ii) There is a splitting of the \(C_{gb}\) curves below \(V_T\) with
drain-to-source bias.

iii) The gate-to-bulk depletion capacitance is reduced.
    This can be seen most clearly from the \(V_{BS} = -2v\)
    curve of the extrinsic plot.

iv) There is a long tail in the depletion-accumulation
    transition (\(V_{GS}\) decreasing below \(V_T\)).

\(C_{gg}\)  
i) The family of drain-to-source bias curves for the
intrinsic part of \(C_{gg}\) (above \(V_T\)) for the short
device lacks the well defined structure of the
saturation-to-linear transitions for the long
device.
ii) There is a splitting of the $C_{gs}$ curves with drain-to-source bias near and below $V_T$.

$C_{bs}$
i) The intrinsic part of $C_{bs}$ (above $V_T$) is greatly reduced for the short-channel device.

ii) There is an appreciable gate bias dependence of the extrinsic capacitance. The variation of the extrinsic capacitance with gate voltage is actually larger than the variation of the intrinsic capacitance. This can be seen from the extrinsic plot.

iii) There is a splitting of the extrinsic capacitance curves with drain-to-source bias.

$C_{bd}$
i) The intrinsic part of $C_{bd}$ (above $V_T$) is greatly reduced for the short-channel device.

ii) There is an appreciable gate bias dependence of the extrinsic capacitance. The variation of the extrinsic capacitance with gate voltage is actually larger than the variation of the intrinsic capacitance.

iii) There is a slight splitting of the extrinsic capacitance curves with drain-to-source bias.

$C_{bg}$
i) The intrinsic part of $C_{bg}$ (above $V_T$) is reduced for the short-channel device.

ii) There is a splitting of the $C_{bg}$ curves below $V_T$ with drain-to-source bias.
iii) There is a long tail to the depletion-accumulation transition ($V_{GS}$ decreasing below $V_T$).

iv) The gate-to-bulk depletion capacitance is reduced for the short device. This can be seen from the $V_{BS} = -2v$ curve of the extrinsic plot.

$C_{bb}$

i) The bulk capacitance above $V_T$ is greatly reduced for the short-channel device.

ii) There is no splitting of the $C_{bb}$ curves below $V_T$ with drain-to-source bias. This indicates that the drain dependence of $C_{bs}$ and $C_{bd}$ is cancelled by the drain dependence of $C_{bg}$.

iii) There is a long tail to the depletion-accumulation transition ($V_{GS}$ decreasing below $V_T$).

iv) The bulk depletion capacitance is reduced for the short device. This can be seen from the $V_{BS} = -2v$ curve of the extrinsic plot.

In addition to these specific observations, there is one characteristic which is common to all of the capacitance curves; for the extrinsic plots, the short-channel device capacitances are always much smaller than the long channel capacitances. This reflects the relatively large importance of the gate-junction overlap capacitances as a fraction of the total gate capacitance. Because of this, the modeling of short-channel capacitances, and particularly the modeling
of the intrinsic parts, is not as critical as the modeling of large-geometry or even narrow-geometry device capacitances. It is important to keep this in mind when the modeling of the short-channel capacitance effects is discussed in the next section.

The most significant short-channel capacitance effects are the reduced substrate capacitances, the bias dependencies of the extrinsic capacitances, and the alterations in the intrinsic channel capacitances due to velocity saturation. The most striking effect of velocity saturation is the non-zero gate-to-drain capacitance in saturation.

Some of these capacitance effects correspond to short-channel effects that can be observed from the transport current. The reduced substrate capacitances are consistent with the reduced body effect factor extracted from the drain current curves. From Table 4.2, the body effect factor, $\gamma$, for a long channel device is 1.04 while $\gamma$ for a short-channel device is only 0.72. This can be seen by the closer grouping of the drain current curves of Fig. A.6a. The smaller body effect is caused by the reduced effective bulk charge along the channel length as the result of charge sharing.

The effect of velocity saturation can be seen by comparing the drain currents of Figs. A.5c and A.6c. For a long channel device, where saturation is dominated by the
Table 4.2

Electrical Parameters of the
20 μm and 4 μm Channel Length Devices

<table>
<thead>
<tr>
<th>Electrical Parameter</th>
<th>20 μm Length</th>
<th>4 μm Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effective Channel Length (V_{GS} ≈ 1V)</td>
<td>17.4 μm</td>
<td>1.4 μm</td>
</tr>
<tr>
<td>Threshold Voltage</td>
<td>0.66 V</td>
<td>0.62 V</td>
</tr>
<tr>
<td>Body Effect Factor (γ)</td>
<td>1.04 V²</td>
<td>0.72 V²</td>
</tr>
<tr>
<td>Mobility Degradation Factor (θ)</td>
<td>0.06 V⁻¹</td>
<td>0.31 V⁻¹</td>
</tr>
</tbody>
</table>

pinch-off constraint, the drain current versus V_{GS} will follow a parabolic curve until the gate voltage becomes large enough to bring the device out of saturation. Beyond this point, the drain current will (asymptotically) increase linearly with V_{GS} (neglecting mobility effects.) For a short-channel device with a large drain-to-source bias, the drain current increases parabolically for low gate voltages, where pinch-off is dominant, and linearly for larger gate voltages (with the device still saturated,) where velocity saturation is dominant. While still in saturation, there is little dependence on the drain voltage.
So, the effects of the physical mechanisms present in short-channel devices can be observed from both the capacitance and static current characteristics. In the next section, simple models for the bulk capacitances and for the gate dependence of the extrinsic capacitances are presented. Also, the expected behavior of the gate-to-channel capacitances for velocity saturated devices is considered, and the observed short-channel capacitance effects are qualitatively discussed.
4.2.3 Modeling short-channel capacitances

The first step in modeling short-channel device capacitances is again the choice of an appropriate threshold voltage model. As indicated above, many short-channel threshold voltage models have been developed using a variety of methods of analysis. The choice of a particular model may be influenced by the physical device structure and will certainly be influenced by the ease of implementation of the model for computer-based calculations. Given a threshold voltage model, the effective body effect factor can be determined, which may be bias dependent as discussed in section 4.1.3. The effective body effect factor for a short-channel device should probably include some dependence on the drain voltage as well as the substrate potential.

The intrinsic bulk capacitances, including $C_{gb}$ and $C_{bg}$ above $V_T$, can be modeled using the large-geometry capacitance expressions of Meyer [4.13] or Ward [4.14] with the effective body effect factor. Although the shapes of the short-channel bulk capacitance curves differ somewhat from the large-geometry bulk capacitance curves, the intrinsic bulk capacitances are so small compared to the fixed junction capacitances for a short-channel device that modeling them is not critical. This can be seen from Fig. 4.14 where the layout of a 5 $\mu$m channel length device (using the 5 $\mu$m design rules of the CMOS process) is shown. The area of the
source and drain junctions is approximately ten times the intrinsic gate area, neglecting width effects. Therefore, the intrinsic bulk capacitances, which are reduced by the short-channel effect, can almost be neglected. Even the bias dependence of the extrinsic bulk capacitances is negligible compared to the fixed junction capacitances. The ratio of the extrinsic bulk capacitances to the intrinsic bulk capacitances will not be as large for SOI (silicon-on-insulator) technologies, where there is little junction capacitance associated with the source and drain contact regions. Nevertheless, using a large-geometry model with the reduced body effect for the short-channel bulk capacitances should produce negligible errors even for SOI.
The next consideration is that of the bias dependencies of the extrinsic capacitances. These can be summarized as follows. For $C_{GS}$ and $C_{GD}$, the extrinsic capacitance decreases as the gate voltage is reduced below $V_T$. For $C_{BS}$ and $C_{BD}$, the opposite occurs; the extrinsic capacitances increase as the gate voltage is reduced below $V_T$. Although the quality of the experimental bulk capacitance data is not very high, these two effects seem to be approximately complementary. For $C_{GB}$ and $C_{BG}$, the principal effect is a splitting of the capacitance curves below $V_T$ with an applied drain-to-source voltage. The physical mechanisms behind these variations in the extrinsic capacitances are considered below, and simple models for these phenomena are proposed.

The bias dependence of the extrinsic gate-to-source (or gate-to-drain) capacitance can be qualitatively understood by considering the field lines of Fig. 4.15. In Fig. 4.15a the field lines in the depletion region are shown for a gate bias just below threshold, and in Fig. 4.15b the field lines are shown at flat-band, which occurs for $V_{GB} - V_{FB} = 0$. The field lines are sketched by neglecting the bulk charge and considering only the potential distribution at the boundaries. Near threshold, the potential variation along the silicon surface is very gentle and it is easy for field lines from the source to circle around and reach the gate. However, near the flat-band condition, the potential varia-
Figure 4.15 – A qualitative sketch of the field lines near the source of a device with idealized junctions: a.) near threshold, and b.) at flat-band.
tion near the surface is comparatively steep. Therefore, the field lines originating at the source have very little vertical component, and there is little coupling from source-to-gate. This conceptual analysis does not consider coupling from the gate to the source through the bulk charge; however, it does qualitatively explain the noted reduction of the gate-to-source fringing capacitance as the surface potential (or the band bending) is reduced with decreasing gate voltage.

In order to correctly calculate the gate-to-source capacitance versus gate voltage, a two-dimensional analysis of the fields in the oxide and the depletion regions must be performed. This is a fundamentally difficult problem because the spatial boundary of the depletion region changes with variations in the terminal voltages. Therefore, gross assumptions must be made about the fields or the charges on the boundaries to obtain a simple analytic solution. The simplest approach is to extend the simple charge analysis of Fig. 4.13 and (4.46). If we assume that one half of the charge in the rectangular area defined by the overlap of the junction and surface depletion regions will be mirrored on the gate, the gate charge will be given by

\[ \Delta Q_G = \frac{1}{2} q N A \bar{W} W \Delta \]

(4.49)
where \( W_j \) is the junction depletion depth given by

\[
W_j = \left[ \frac{2 \varepsilon_{Si}}{qN_A} (V_{SB} + \phi_B) \right]^{\frac{1}{2}}
\]

or

\[
W_j = \left[ \frac{2 \varepsilon_{Si}}{qN_A} (V_{DB} + \phi_B) \right]^{\frac{1}{2}} \tag{4.50}
\]

and where \( W_d \) is the depletion depth under the gate which is given by

\[
W_d = \left[ \frac{2 \varepsilon_{Si}}{qN_A} \psi_s \right]^{\frac{1}{2}} \tag{4.51}
\]

Recall that \( \psi_s \) is the electrostatic potential at the silicon surface relative to the potential deep in the bulk, and \( \phi_B \) is the built-in potential of the source and drain junctions. Incorporating (4.50) and (4.51) into (4.49) produces

\[
\Delta Q_G = \varepsilon_{Si} \left[ (V_{SB} + \phi_B) \psi_s \right]^{\frac{1}{2}}
\]

and

\[
\Delta Q_G = \varepsilon_{Si} \left[ (V_{DB} + \phi_B) \psi_s \right]^{\frac{1}{2}} \tag{4.52}
\]

for the source end of the channel and the drain end of the channel respectively. Taking the partial derivative of (4.52) with respect to \( V_S \) (or \( V_D \)) produces the excess gate-to-source and gate-to-drain capacitances per unit width. With some rearranging of the leading coefficient, the normalized capacitances can be written as
\[
\frac{\Delta C_{gs}}{L C_{ox}} = \frac{1}{2} \frac{\varepsilon_{si}}{\varepsilon_{ox}} \frac{t_{ox}}{L} \left[ \frac{\psi_s}{(V_{SB} + \phi_B)} \right]^3
\]

and

\[
\frac{\Delta C_{gd}}{L C_{ox}} = \frac{1}{2} \frac{\varepsilon_{si}}{\varepsilon_{ox}} \frac{t_{ox}}{L} \left[ \frac{\psi_s}{(V_{DB} + \phi_B)} \right]^3.
\]

These expressions agree very well qualitatively and fairly well quantitatively with the experimental data. In particular, for \(\Delta C_{gs}\), (4.53a) reaches a maximum as \(\psi_s\) approaches its maximum value of \(V_{SB} + 2\phi_F\) at threshold and equals zero when \(\psi_s\) goes to zero at flat-band. At threshold, the square-root term in (4.53a) is approximately equal to 1 since \(\phi_B \approx 2\phi_F\). (Actually, \(\phi_B\) is larger than \(2\phi_F\) by about 150 mV.) So, the change in \(C_{gs}\) from flat-band to threshold should be approximately 0.085 \(C_{oi}\) given the oxide thickness and the intrinsic length of the n-channel transistor. This agrees fairly well with the actual variation of the extrinsic part of \(C_{gs}\) in Fig. A.6e. (The flat-band voltage is estimated to be 1.8V below threshold.)

Qualitatively, if a source-to-bulk bias is applied, the value of (4.53a) at threshold remains fixed but the excess capacitance decreases more slowly with changes in the gate voltage. The excess capacitance will continue to decrease with \(V_{GS}\) until (4.53a) again reaches zero at the new gate flat-band voltage, which tracks the source-to-bulk bias. If a drain-to-source bias is applied, the denominator in the
square-root term of (4.53b) is increased but $\psi_s$ is unaffected. Therefore, near threshold, $\Delta C_{gd}$ is larger for a lower drain-to-source voltage, but at flat-band, $\Delta C_{gd}$ is zero regardless of the drain-to-source bias. Notice in Fig. A.6h that the $V_{DS} = 0\text{v}$ curves lie above the $V_{DS} = 2\text{v}$ curves until they meet approximately at flat-band. Also notice that the values of $C_{gde}$ at flat-band are approximately equal independent of the source-to-bulk bias.

Unfortunately, (4.53) only applies when the silicon surface is depleted, i.e., when $\psi_s > 0$ and $V_{GB} < V_{PB}$. If the gate voltage is reduced below the gate flat-band voltage, holes (positive charge) will accumulate at the surface and the surface potential will be negative. The magnitude of the surface potential will increase approximately logarithmically with the falling gate voltage. This suggests that an empirical model with a logarithmic dependence on the gate voltage could be used for the excess gate-to-source and gate-to-drain capacitances in accumulation. The importance of these capacitances should be kept in perspective, however; the total variation of the extrinsic capacitances over the complete range of gate voltage is only about 8% of $C_{oe}$. A capacitance variation of this size is probably negligible for most circuit applications.

For the bulk-to-source and bulk-to-drain capacitances, the excess extrinsic capacitances can be modeled by assuming
that the increase in these capacitances below threshold is equal to the decrease in the gate-to-source and gate-to-bulk extrinsic capacitances. This assumption seems reasonable since the total bulk charge that will be mirrored on the source (or drain) is equal to the total charge in the junction depletion region less the gate's share of the charge in the overlap. From Figs. A.6n and A.6p, it seems that the variation in the extrinsic bulk capacitances is somewhat larger than the corresponding variations in the extrinsic gate capacitances. Moreover, there seems to be a significant variation in $C_{bse}$ with drain voltage even though the drain should be isolated from the source junction. This probably indicates that the rectangular approximation to the depletion region of Fig. 13 is not accurate. This should not be surprising considering that the depletion regions are on the order of 0.35 μm wide, which is an appreciable fraction of the effective channel length. In spite of these discrepancies, the suggested model for the extrinsic bulk capacitances ($C_{bs}$ and $C_{bd}$) should certainly be accurate enough for most applications, particularly considering the overwhelming significance of the fixed junction capacitances.

Concluding the discussion of the short-channel extrinsic capacitances, the drain voltage dependence of the gate-to-bulk and bulk-to-gate capacitances below threshold is considered. These capacitances are associated with the
simple MOS capacitor between the gate and the bulk, and for a large-geometry device, the source and drain have no effect on these capacitances. To first order, the effective area of the MOS capacitor is given by the device width multiplied by the distance between the source and drain junctions. So, the short-channel gate-to-bulk and bulk-to-gate capacitances can be modeled by the capacitances expected for a large-geometry device scaled by the effective channel length. In other words,

\[ C_{gb} = \frac{C_{gb}}{L} \left( L - W_{js} - W_{jd} \right) \]  
\[ C_{bg} = \frac{C_{bg}}{L} \left( L - W_{js} - W_{jd} \right) \]  

where \( W_{js} \) and \( W_{jd} \) are the junction depletion widths for the source and drain respectively, and \( C_{gb} \) and \( C_{bg} \) are the capacitances expected for a large-geometry device of length \( L \) (the electrical channel length.) The splitting of the experimental \( C_{gb} \) and \( C_{bg} \) curves is explained by the reduced effective length with an applied drain-to-source bias. Eq. 4.54 also predicts that the accumulation capacitances will be reduced with substrate bias due to the increase of both \( W_{js} \) and \( W_{jd} \). This can be observed in Figs. A.6.j and A.6.r. Equation 4.54 will only apply as long as the source and drain junction regions stay far enough apart to justify the rectangular approximation to the depletion regions of
Fig. 4.13. Relative to the experimental data, (4.54) over-estimates the reduction of the effective length, which implies that this approximation may not be valid. An estimate of the charge sharing in the two-dimensional depletion regions is required to calculate the effective channel length for these capacitances when the rectangular approximation is inappropriate.

Finally, the effect of velocity saturation on the device capacitances is considered. First, the static transport current of a velocity saturated device is calculated using a simple model for the velocity-field relationship. This gives some insight into the velocity-limited saturation voltage and the distribution of charge in the channel.

It will be assumed that the velocity-field relationship can be described by a constant mobility for fields up to $E_C$ and a constant saturation velocity $v_S$ for fields greater than $E_C$. Furthermore, it will be assumed that the device is saturated and that the saturation velocity is reached at $y=f$, which is also the edge of the drain region. With these assumptions, the current transport equation at the drain can be written as

$$-I_d = Wv_S Q_n(f) . \quad (4.55)$$

If bulk charge is ignored and the threshold voltage is constant along the channel, $Q_n(f)$ can be written as
\[ Q_n(l) = -C_{ox}(V_{gs} - V_{DSSat}) \]  

(4.56)

where \( V_g = V_G - V_T \). For \( V_{DS} = V_{DSSat} \), the channel is not pinched off, so the drain current can be written as

\[ I_d = \frac{\mu_{WC_{ox}}}{l}[V_{gs}V_{DSSat} - \frac{1}{2}V_{DSSat}^2] \]  .

(4.57)

Equating (4.55) and (4.57), \( V_{DSSat} \) can be calculated as

\[ V_{DSSat} = (V_{gs} + E_{C}l) - [V_{gs}^2 + (E_{C}l)^2]^{\frac{1}{2}} \]  .

(4.58)

where \( E_c = V_S/\mu \). If \( E_c \) is very large, \( V_{DSSat} \) approaches \( V_{gs} - V_T \) as expected, and if \( E_c \) is very small, \( V_{DSSat} \) approaches \( E_{C}l \). From (4.55), the drain current under velocity saturation can be calculated as

\[ I_d = WC_{ox}V_S\left[(E_c l)^2 + V_{gs}^2\right] - E_{C}l \]  .

(4.59)

To calculate the device capacitances, the channel and drain regions must be considered separately. In the channel region, the Meyer [4.13] or Ward [4.14] methods can be used to calculate the device capacitances by taking the partial derivatives of the total gate, bulk, and channel charges. The results will differ from the results for large-geometry devices because of the changes in \( V_{DSSat} \) and because of the dependence of \( l \) on \( V_{DS} - V_{DSSat} \). In actuality, these expressions are very complicated, particularly if a gentler
velocity-field relationship is used. Much of the mobile charge in the drain region will be mirrored on the drain rather than on the gate. This is the case because the lateral field and the divergence of the lateral field will be larger than the vertical field and the divergence of the vertical field throughout most of the drain region. In order to calculate the terminal capacitances, a two-dimensional partitioning of charges in the drain region must be performed, possibly using a method similar to that used for the short- and narrow-channel threshold voltages. In this case, however, the distribution of the carriers in the drain region channel is much more complicated than the distribution of fixed charge in the non-saturated case.

As a result of these difficulties, a thorough analysis of the short-channel capacitances was not attempted. However, some simple qualitative analyses can be used to explain the observed short-channel behavior. First, the increased gate-to-source capacitance in saturation can be understood by considering the distribution of channel charge for the pinch-off and the velocity-saturation cases.

Fig. 4.16 shows $-Q_n(y)$ for two devices; Fig. 4.16a is a long channel device where saturation is controlled by pinch-off, and Fig. 4.16b is a short-channel device exhibiting velocity-limited saturation. In each case, the channel charge is shown for two different values of $V_{GS}$. In the
Figure 4.16 - A qualitative sketch of the mobile channel charge versus position for a.) a large-geometry device, and b.) a short-channel device exhibiting velocity saturation.
first case, $Q_n(l)$ is pinned at zero, and as the gate-to-source voltage is increased the whole curve scales without changing shape. Therefore, $C_{GS}$ is independent of $V_{GS}$, and the value of the capacitance is determined by the shape of the charge distribution. In the second case, the value of $Q_n(l)$ changes as well as $Q_n(0)$. The gate-to-source capacitance is bias dependent, since the shape of the charge distribution is changing as a function of $V_{GS}$, and the capacitance is larger than before because there is more modulation of the charge near the drain. Also, from (4.55), $Q_n(0)$ is proportional to $I_d$. Therefore, since $dI_d/dV_{GS}$ is always increasing with $V_{GS}$, $C_{GS}$ will increase with $V_{GS}$ until the device comes out of saturation. This is also observed from the experimental data.

For $C_{GD}$, there are two mechanisms which can cause there to be a non-zero gate-to-drain capacitance in saturation. First, there is the modulation of the channel length with the voltage drop across the drain region. This introduces a variation in the total channel charges of Fig. 4.16 which are mirrored on the gate. Therefore, there is a component of $dQ_G/dV_D$ which arises through $dI/dV_D$. Second, there is the mobile charge in the drain region which will be partially mirrored on the gate. The total charge in the drain region is determined by the channel charge density at the edge of the drain region, $Q_n(l)$, and the width of the drain region, $AL = L-l$. A gate-to-drain capacitance will be
produced in this case by the modulation of the drain region width with drain voltage, and possibly by a drain bias dependence of the charge sharing. The first mechanism will dominate when \( Q_n(f) \) is small (or the transport current is small,) and the second mechanism will dominate when \( Q_n(f) \) (or the transport current is large.) In either case, however, the gate-to-drain capacitance will be driven by the modulation of the drain region width, which is very difficult to calculate.

As a final note to short-channel capacitance modeling, it is worth mentioning the agreements and disagreements of the experimental data with the theoretical short-channel model of Taylor et al [4.41]. The Taylor model does predict the non-zero gate-to-drain capacitance in saturation. This theory ignores any contribution of the drain region charge to the gate charge, so this capacitance is predicted solely on the basis of the impact of the modulation of the channel length on the channel region charge. Unfortunately, the data presented in [4.41] is not sufficient to determine whether the model predicts the observed increase of \( C_{GS} \) in saturation. (The gate-to-source capacitance is only plotted for a single, low drain-to-source bias.) However, this theory predicts one phenomenon that was not observed. Taylor's data shows \( C_{gd} \) for low drain voltages dropping below 0.5 at high gate voltages. No explanation for this phenomenon is given, and it certainly does not seem to agree
with simple intuition. It should also be noted that there are problems with Taylor's treatment of the channel capacitances \(C_{sg}, C_{dg}, C_{sb},\) and \(C_{db}\). He uses an ad hoc channel partitioning scheme, as discussed in section 2.3.2, and the plotted values for \(C_{sg}\) differ with Ward's theory and data by a factor of two.

As discussed above and in Appendix A, the capacitances of small-geometry devices are heavily dominated by the fixed extrinsic capacitances and, ultimately, the circuit wiring and load capacitances. In view of this, accurate small-geometry device capacitance models may not be important for most circuit applications. However, it is important to have some understanding of the divergences from simple theory of short- and narrow-channel device behavior. The models discussed here are presented in this spirit, and although these models are not fully developed, they are consistent with the trends observed in the experimental data.
CONCLUSIONS

Summary

The principal goal of this work has been the development of a technique for the measurement of small-geometry MOS transistor capacitances. Capacitance data from individual, small-geometry devices are crucial to the modeling and characterization of the short- and narrow-channel capacitance effects found in modern MOSFET's. An integrated circuit technique has been described that uses on-chip circuits to sense small capacitive currents, producing a low-impedance signal that can be interpreted off-chip using a conventional gain-phase instrument. The closed-loop coulombmeter technique has been demonstrated in a 5 μm CMOS process, and gate and bulk capacitance data have been obtained from devices as small as 4 μm by 4 μm. The technique is scalable, and will allow the measurement of minimum-geometry devices in state-of-the-art MOS technologies of the future.

The on-chip capacitance measurement technique has two principal limitations. First, the technique can only be used to measure low-leakage capacitors. Therefore, this approach can not be used to measure capacitances at the source and drain, or bulk capacitances for devices with
large hot-electron currents. Second, the technique requires a fairly mature process to support the design of a unity-gain stable operational amplifier. Nevertheless, the closed-loop coulombmeter technique should prove to be a useful technique for obtaining small-geometry capacitance data for both modeling and characterization purposes.

A secondary goal of this work has been to use the capacitance data described above to develop models for short- and narrow-channel capacitance effects. Detailed models for narrow-channel capacitance effects have been described which are in good agreement with the experimental data. These models predict the increase of the gate and bulk capacitances as the device channel width increases with increasing gate voltage. For short-channel devices, there are large bias-dependencies of the extrinsic capacitances, i.e., the capacitances associated with the gate-to-source and gate-to-drain overlaps and the source and drain junctions. A simple model for these effects has been described which qualitatively predicts the observed phenomena. The other major short-channel capacitance effects are related to velocity saturation. These effects have been discussed, however, detailed models have not been proposed. The velocity saturation effects observed for the 4 μm (1.4 μm effective) channel length devices studied here were relatively minor. Data from shorter devices should be used in the future to investigate these effects further.
In addition to this work on the measurement and modeling of small-geometry device capacitances, the theory of large-geometry MOSFET capacitance modeling has been reviewed. The issues that have been discussed here are: the importance of using large-signal models for transient analysis to avoid charge-conservation problems, the methods used by several authors to partition the channel charge into individual source and drain components, and the conservation-of-energy problem for non-reciprocal capacitance models. Also, the high-frequency behavior of the reciprocal Meyer model and the non-reciprocal Ward model have been compared to a non-quasi-static solution for the MOSFET small-signal admittances. The non-quasi-static result provides some insight into the origin of the non-reciprocal capacitances of the Ward model, and suggests an improved high-frequency model based on a reciprocal admittance matrix.

Recommendations for Further Work

With respect to capacitance measurements, there are three areas that warrant further work. First, an implementation of the closed-loop coulombmeter technique in a micron or submicron process would yield better data for studying velocity-saturation related short-channel effects, and would demonstrate the scalability of the on-chip approach to capacitance measurements. Second, a synchronous method for
making the gain-phase measurement could be developed which would completely eliminate the dominant source of noise for the measurements made here, namely, the interference of the pulses produced by the reset switch. This requires an instrument with both good noise rejection and a fast acquisition time, which are generally mutually exclusive. Finally, the quasi-bridge technique, discussed in section 3.2.3, should be explored as a vehicle for making source and drain capacitance measurements on-chip. In addition to allowing the measurement of capacitances in the presence of large currents, the quasi-bridge technique does not require sensitive on-chip circuitry, as does the coulombmeter technique. Therefore, this approach is potentially suited to a broader range of measurement applications.

In the area of small-geometry capacitance modeling, the models developed here for narrow-channel effects are fairly complete; however, they could be improved by using a better approximation to the shape of the oxide in the transition region. Also, a more careful consideration of the differing effects of surface states (oxide charge) and threshold tailoring implants should be made. More fundamentally, the analysis presented only develops the narrow-channel gate and bulk capacitances, rather than the gate and bulk charges. For charge-based, large-signal analysis, a charge description is preferred, but this requires a more complete analysis of the potential and the charges along the oxide taper.
For short-channel effects, a complete analysis of the various charge components along the channel should be performed. The critical issue for capacitance modeling is the distribution of charge in the drain region, and the way that this charge is mirrored at the gate and at the drain. In addition, the coupling between the spatial modulation of the drain region and the total channel region charge must be considered. For both short- and narrow-channel capacitance modeling, it will be important that good experimental data from several technologies are available so that the small-geometry effects can be correlated to differences in the physical device structures.

Finally, although the non-quasi-static analysis of Section 2.4 suggests an improved high-frequency model for the MOSFET, the analysis presented here is not complete for two reasons. First, the non-quasi-static solution of Section 2.4 is strictly a small-signal model. As such, it is not directly applicable to charge-based, large-signal transient analysis. Second, a solution was only obtained for the case of zero bulk charge. Since the non-quasi-static treatment of the MOSFET is equivalent to that of a nonlinear transmission line, a closed form solution is not guaranteed for the general case. Therefore, a practical model based on this analysis may require an ad hoc extrapolation of the model presented here. Such an extrapolation could be confirmed by numerical analysis, and, if such a
model could be developed, it would be more accurate than the Meyer or Ward quasi-static models for high-frequency circuit simulation.
Appendix A

Gate and Bulk Capacitance Data

The gate and bulk capacitance data obtained using the CMOS coulombmeter circuits are presented at the end of this Appendix. The data are organized according to the test device geometries, which are (in order): 100/100, 20/20, 20/4, 4/20, and 4/4 W(μm)/L(μm). For each geometry, three plots of static current and one plot of DC transconductance are included, all of which were taken from the gate coulombmeter test device. The four gate capacitances (C_{gs}, C_{gd}, C_{gb}, and C_{gg} respectively) and the four bulk capacitances (C_{bs}, C_{bd}, C_{bg}, and C_{bt} respectively) follow, with each capacitance plotted for two bias sweeps. The two bias sweeps, and the normalization of the data for these two plots, have been chosen to facilitate the direct comparison of both the intrinsic and extrinsic capacitances of different device geometries. The terms intrinsic and extrinsic as used here are carefully defined below, and the normalization calculations are described in detail.
A.1 Intrinsic and extrinsic device capacitances

The terminal capacitances of real MOS transistors are the sum of two conceptually distinct terms: the intrinsic capacitances, which come from the channel region of the device, and the extrinsic capacitances which are produced by the non-idealities of the physical device structure. The intrinsic capacitances are primarily associated with the inversion channel charge and were discussed in Chapter 2. The extrinsic capacitances come from the source and drain junctions, and the lateral diffusion of the source and drain junctions under the gate which produce the gate-to-source and gate-to-drain overlaps. The terms intrinsic and extrinsic as used here refer to the physical sources of the two components of the device capacitances, the intrinsic device (the channel) and the extrinsic device (the junctions and overlaps.)

For experimental capacitance data there is a third component of the measured capacitances. Stray capacitances are introduced by the unavoidable coupling between the physical substrate and the metal and poly traces needed to connect to the source, drain, and gate. The sources of these stray capacitances can be seen from the idealized device layout of Fig. A.1. The metal traces to the gate, source, and drain contacts (dashed lines) must cross over the p-well of the device (the surrounding rectangle)
Figure A.1 - An idealized layout of an n-channel transistor in the p-well.
producing a stray capacitance to the substrate. In addition, there is stray capacitance between the device well and the polysilicon of the gate under both the gate contact area and the polysilicon overlap over the field oxide. These stray capacitances are independent of the measurement technique since contact must be made to the four device terminals regardless of the measurement method.

The total composition of the measured gate and bulk capacitances is summarized in (A.1).

\[
\begin{align*}
C_{gs} &= C_{gsi} + C_{gse} \\
C_{gd} &= C_{gdi} + C_{gde} \\
C_{gb} &= C_{gbi} + C_{gbe} + C_s \\
C_{bs} &= C_{bsi} + C_{bse} + C_s \\
C_{bd} &= C_{bdi} + C_{bde} + C_s \\
C_{bg} &= C_{bgi} + C_{bge} + C_s
\end{align*}
\]  

(A.1)

This mix of intrinsic, extrinsic, and stray capacitances in the measured data makes it impossible to compare raw data from different device geometries directly. In particular, the relative sizes of the intrinsic, extrinsic, and stray capacitances change as the device geometry changes. The extrinsic and stray capacitances become much more significant for small-geometry devices. For example, the gate-to-source and gate-to-drain overlap capacitances are the same.
for a 4 μm long device and a 100 μm long device; however, the intrinsic gate-to-source and gate-to-drain capacitances scale by a factor of more than 50. Therefore, some data reduction is required to separate, at least roughly, the various capacitance components to facilitate direct comparisons. Fortunately, the intrinsic, extrinsic, and stray capacitances have different bias dependencies which can be exploited for this purpose.

The intrinsic capacitances can easily be extracted from the total measured capacitances because of their dependence on gate voltage. In particular, the intrinsic capacitances associated with the channel charge \( C_{g_{si}}, C_{g_{di}}, C_{b_{si}}, \) and \( C_{b_{di}} \) are zero when the device is off and the channel charge is zero. However, the associated extrinsic and stray capacitances are roughly independent of gate voltage. Therefore, the intrinsic capacitances are simply the added capacitance observed as the gate voltage is swept from below the threshold voltage \( (V_T) \) to above the threshold voltage. The gate-to-bulk and bulk-to-gate capacitances, on the other hand, have exactly the opposite relationship: \( C_{g_{bi}} \) and \( C_{b_{gi}} \) are approximately zero when the gate voltage is very high and \( V_{DS} = 0 \). Therefore, the intrinsic parts of \( C_{gb} \) and \( C_{bg} \) are simply the added capacitance observed as device turns off.
Separating the remaining extrinsic and stray capacitances is somewhat more difficult. For $C_{gs}$ and $C_{gd}$ there is no problem; the extrinsic parts of $C_{gs}$ and $C_{gd}$ come from the approximately bias-independent overlap capacitances, and there are no stray capacitances, as discussed above. The extrinsic parts of $C_{bs}$ and $C_{bd}$ come from the source-to-bulk and drain-to-bulk junction capacitances. These junction capacitances vary with the source-to-bulk and drain-to-bulk potentials according to the power law and the built-in potential for the particular doping profile. The stray parts of $C_{bs}$ and $C_{bd}$, on the other hand, are completely bias independent. Therefore, the extrinsic and stray parts of these capacitances can be separated by fitting the data to the junction capacitance characteristics and taking the offset as the stray capacitance.

Physically, the extrinsic parts of $C_{gb}$ and $C_{gb}$ are associated with the lateral growth of the field oxide which encroaches upon the drawn gate width. As discussed in Chapter 4, this extrinsic capacitance has a strong bias dependence on $V_{GS}$ for narrow channel devices. This bias dependence causes problems with the intrinsic capacitance extraction outlined above. Moreover, the fixed parts of $C_{gbe}$ and $C_{bge}$ are indistinguishable from the stray capacitances associated with the gate polysilicon and metal traces. Therefore, for $C_{gb}$ and $C_{bg}$, it is impossible to separate neatly the intrinsic, extrinsic, and stray parts.
from the raw data alone. Instead, it is necessary to evaluate the stray capacitances separately using other test structures, so that they can be subtracted from the data for $C_{gb}$ and $C_{bg}$. This requires an accurate absolute scale factor in the coulombmeter circuit which, unfortunately, was not achieved here. As a result, in the normalization calculations described below no attempt was made to isolate the stray and extrinsic parts of $C_{gb}$ and $C_{bg}$ capacitances individually.

This problem with $C_{gb}$ and $C_{bg}$ points out the fact that the distinction between the intrinsic, extrinsic, and stray capacitances is somewhat artificial, particularly for small geometries where there are appreciable interactions between the intrinsic and extrinsic structures of the device. Nevertheless, it is important to separate, at least grossly, these contributions in order to compare data from devices with different geometries. The normalization calculations described in the following section were used to isolate the intrinsic and extrinsic capacitance phenomena in the capacitance plots which follow.
A.2 The normalization calculations

The gate and bulk capacitance data have been normalized and plotted for two different bias sweeps. The first plot shows capacitance versus $V_{GS}$ with $V_{DS}$ as a parameter. The gate-to-source voltage is swept from 2 volts below $V_T$ (the actual measured threshold voltage for the device) to 6 volts above $V_T$. The drain-to-source voltage is equal to 0v, 1v, 2v, and 3v for the four parametric curves. This plot is intended to show the behavior of the intrinsic capacitances when the device is on. The second plot is of capacitance versus $V_{GS}$ with both $V_{DS}$ and $V_{BS}$ as parameters. In this case, $V_{GS}$ varies from 6v below threshold to 2v above. The solid curves are for $V_{BS}$ equal to 0v with $V_{DS}$ equal to 0v and 2v, and the broken lines are for $V_{BS}$ equal to -2v with $V_{DS}$ again equal to 0v and 2v. This second plot is intended to show the behavior of the extrinsic device capacitances when the device is off.

The data for the first (intrinsic) plot are the intrinsic capacitances normalized by the measured intrinsic oxide capacitance. The intrinsic capacitances are calculated by subtracting from the total capacitances a fixed estimate of the extrinsic and stray capacitances. These calculations are summarized below.
\[ C_{gsi} = C_{gs} - C_{gs} \]
\[ V_{GS} - V_T = -2 \text{v}, \; V_{DS} = 0 \text{v} \]
\[ C_{gdi} = C_{gd} - C_{gd} \]
\[ V_{GS} - V_T = -2 \text{v}, \; V_{DS} = 0 \text{v} \]
\[ C_{gbi} = C_{gb} - C_{gb} \]
\[ V_{GS} - V_T = 6 \text{v}, \; V_{DS} = 0 \text{v} \]
\[ C_{ggi} = C_{gsi} + C_{gdi} + C_{gbi} \]
\[ C_{bsi} = C_{bs} - C_{bs} \]
\[ V_{GS} - V_T = -2 \text{v}, \; V_{DS} = 0 \text{v} \]
\[ C_{bdi} = C_{bd} - C_{bd} \]
\[ V_{GS} - V_T = -2 \text{v}, \; V_{DS} = V_{DS} \]
\[ C_{bgi} = C_{bg} - C_{bg} \]
\[ V_{GS} - V_T = 6 \text{v}, \; V_{DS} = 0 \text{v} \]
\[ C_{bgi} = C_{bsi} + C_{bdi} + C_{bgi} \] (A.2)

For \( C_{gs} \) and \( C_{gd} \), which have no stray component, the extrinsic capacitances are taken as the value of \( C_{gs} \) and \( C_{gd} \) at \( V_{GS} - V_T = -2 \text{v} \). This gate bias corresponds roughly to the flat-band voltage and is conveniently at the far left of the plot. The extrinsic plus stray capacitances for \( C_{bs} \) and \( C_{bd} \) are treated the same way, except that for \( C_{bd} \) a separate value is needed at each drain voltage in order to accommodate the drain-to-bulk voltage dependence of the extrinsic junction capacitance. For \( C_{gb} \) and \( C_{bg} \) the extrinsic plus stray capacitances are taken as the values of \( C_{gb} \) and \( C_{bg} \) at \( V_{GS} - V_T = 6 \text{v} \), which is the highest gate voltage available, and \( V_{DS} = 0 \text{v} \). As a result of these normalizations, the
intrinsic capacitance curves always have \( c_{gs}, c_{gd}, c_{bs}, \) and \( c_{bd} \) equal to zero at the far left, and \( c_{gb} \) and \( c_{bg} \) equal to zero the at the far right.

The normalizing scale factor for the intrinsic plots is the intrinsic oxide capacitance, \( c_{oi} \), which is defined by

\[
c_{oi} = c_{gsi} + c_{gdi} \quad \text{for } V_{GS} - V_T = 6V, V_{DS} = 0v
\]  \hspace{1cm} (A.3)

The intrinsic oxide capacitance corresponds to the oxide capacitance defined by the intrinsic device length and the intrinsic device width. This is shown schematically in Fig. A.2. Using this normalization factor, the intrinsic plots of \( c_{gs} \) and \( c_{gd} \) for \( V_{DS} = 0v \) are approximately equal to

![Diagram of intrinsic capacitance](image)

**Figure A.2 - A schematic representation of \( c_{oi}. \)**
0.5 at the far right. The effectiveness of this normalization procedure can be gauged by the ease with which the resulting intrinsic capacitance plots for different device geometries can be compared.

The intrinsic capacitances alone, however, do not give a complete picture of the total device capacitances. The second (extrinsic) plot for each of the device capacitances is provided to demonstrate the voltage dependencies of the extrinsic capacitances and the relative importance of the intrinsic and extrinsic capacitances for the different device geometries. The capacitances shown in the extrinsic plots are not the extracted extrinsic capacitances as described in the previous section. Rather, the sweep of $V_{GS}$ has been chosen to highlight the variations in the plotted capacitances which are attributable to the extrinsic device with $V_{GS}$ less than $V_T$. For $C_{gs}$ and $C_{gd}$ the total measured capacitance, intrinsic plus extrinsic, is used for the extrinsic plots. For $C_{gb}$, $C_{bg}$, $C_{bs}$, and $C_{bd}$ the intrinsic capacitances as calculated above are used. This was necessary to eliminate the large stray capacitances in $C_{gb}$ and $C_{bg}$, and the large junction capacitances in $C_{bs}$ and $C_{bd}$ which vary with the source-to-bulk and drain-to-bulk potentials. These large fixed components would have made it impossible to show these capacitances on a uniform scale. (As above, $C_{gg}$ and $C_{bb}$ are defined as the sum of the other three gate or bulk capacitances.)
The normalization factor for the extrinsic plots is the extrinsic oxide capacitance defined by

$$C_{oe} = C_{gs} + C_{gd} \quad \text{for } V_{CS} - V_T = 6v, \; V_{DS} = 0v$$  \quad (A.4)

The extrinsic oxide capacitance, $C_{oe}$, corresponds to the oxide capacitance defined by the total drawn device length and the intrinsic device width. This is shown schematically in Fig. A.3. This value is very close to the total gate capacitance, but is still substantially smaller than the capacitance predicted by the drawn device dimensions when the channel is narrow.

![Figure A.3 - A schematic representation of $C_{oe}$](image-url)
(The calculation of $C_{oe}$ and $C_{oi}$ above, can only be carried out for the gate coulombmeter. Since the scale factor for the bulk coulombmeter was not the same as for the gate coulombmeter, $C_{oi}$ and $C_{oe}$ for the bulk capacitances were calculated by scaling their gate coulombmeter values by the ratio of $C_{bg\;i}$ to $C_{gb\;i}$ for $V_{GS}$ equal to $-6\text{v}$. These capacitances are reciprocal for this bias condition and should therefore be equal.)

The importance of the extrinsic capacitances is demonstrated in Table A.1. The ratio of $C_{oi}$ to $C_{oe}$ is approximately equal to the ratio of the intrinsic gate length to the total gate length. For the 4 $\mu$m channel length devices, the extrinsic overlap capacitances are roughly two-thirds of the total gate capacitance. The ratio of $C_{jo}$ to $C_{oe}$, where $C_{jo}$ is the zero bias source/drain junction capacitance, indicates the importance of the extrinsic parts of $C_{bs}$ and $C_{bd}$. For the short and narrow channel devices, the extrinsic source-to-bulk and drain-to-bulk capacitances are several times larger than the intrinsic capacitances. For the 4 $\mu$m long device, with the source and drain areas determined by 5 $\mu$m design rules, the source and drain junction capacitances are actually larger than the total gate capacitance. The last entry in the table is the ratio of $C_{gb\;e}$ (or $C_{bge}$) to $C_{oe}$, which indicates the importance of simple wiring capacitances for small geometry devices. For the smaller devices, the gate overlap and gate
Table A.1

Measures of the Importance of the Extrinsic Capacitances of Small-Geometry Devices

<table>
<thead>
<tr>
<th>Device Geometry</th>
<th>$\frac{C_{oi}}{C_{oe}}$</th>
<th>$\frac{C_{jo}}{C_{oe}}$</th>
<th>$\frac{C_{gbe}}{C_{oe}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 x 100</td>
<td>0.98</td>
<td>0.09</td>
<td>0.01</td>
</tr>
<tr>
<td>20 x 20</td>
<td>0.89</td>
<td>0.46</td>
<td>0.06</td>
</tr>
<tr>
<td>20 x 4</td>
<td>0.35</td>
<td>3.54</td>
<td>0.15</td>
</tr>
<tr>
<td>4 x 20</td>
<td>0.87</td>
<td>1.44</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Interconnect capacitances are an appreciable fraction of the gate capacitance. The implication of this data is that the modeling of the extrinsic capacitances of small-geometry devices is at least as important as the modeling of the intrinsic capacitances. This can be seen in the extrinsic plots of the gate and bulk capacitance data.
A.3 Final notes on the capacitance data

There are two exceptions to the presentation of the data as outlined above. The first is that the bulk capacitances for the 4 μm by 4 μm device are not included. For this device, the extrinsic junction capacitances are so large compared to the intrinsic bulk capacitances that it was impossible to obtain meaningful data for the intrinsic bulk-to-source and bulk-to-drain capacitances. Second, the drain-to-source bias for the bulk capacitances of the 4 μm channel length device was limited to 2V to avoid hot electron currents in the substrate. (See Chapter 3.)

It should be noted that there is some variation in the scale factors for the plots which follow. Care should be taken when comparing curves to verify that the scale factors are the same.
Figure A.4a - Drain current versus $V_{GS}$ with $V_{BS}$ as a parameter for a 100 $\mu$m by 100 $\mu$m (W/L) device.

Figure A.4b - Drain current versus $V_{DS}$ with $V_{GS}$ as a parameter for a 100 $\mu$m by 100 $\mu$m (W/L) device.
Figure A.4c - Drain current versus $V_{GS}$ with $V_{DS}$ as a parameter for a 100 μm by 100 μm (W/L) device.

Figure A.4d - Transconductance versus $V_{GS}$ with $V_{DS}$ as a parameter for a 100 μm by 100 μm (W/L) device.
Figure A.4e - Intrinsic plot of $C_{gs}$ for a 100 $\mu$m by 100 $\mu$m (W/L) device, normalized by $C_{oi}$.

Figure A.4f - Extrinsic plot of $C_{gs}$ for a 100 $\mu$m by 100 $\mu$m (W/L) device, normalized by $C_{oe}$. 

- 252 -
Figure A.4g - Intrinsic plot of $C_{gd}$ for a 100 μm by 100 μm (W/L) device, normalized by $C_{oi}$.

Figure A.4h - Extrinsic plot of $C_{gd}$ for a 100 μm by 100 μm (W/L) device, normalized by $C_{oe}$. 

- 253 -
Figure A.4i - Intrinsic plot of $C_{gb}$ for a 100 μm by 100 μm (W/L) device, normalized by $C_{oi}$.

Figure A.4j - Extrinsic plot of $C_{gb}$ for a 100 μm by 100 μm (W/L) device, normalized by $C_{oe}$.
Figure A.4k - Intrinsic plot of $C_{gg}$ for a 100 μm by 100 μm (W/L) device, normalized by $C_{oi}$.

Figure A.4l - Extrinsic plot of $C_{gg}$ for a 100 μm by 100 μm (W/L) device, normalized by $C_{oe}$.
Figure A.4m – Intrinsic plot of $C_{bs}$ for a 100 μm by 100 μm (W/L) device, normalized by $C_{oi}$.

Figure A.4n – Extrinsic plot of $C_{bs}$ for a 100 μm by 100 μm (W/L) device, normalized by $C_{oe}$.
Figure A.4o - Intrinsic plot of $C_{bd}$ for a 100 μm by 100 μm (W/L) device, normalized by $C_{oi}$.

Figure A.4p - Extrinsic plot of $C_{bd}$ for a 100 μm by 100 μm (W/L) device, normalized by $C_{oe}$. 

- 257 -
Figure A.4q - Intrinsic plot of $C_{bg}$ for a 100 μm by 100 μm (W/L) device, normalized by $C_{oi}$.

Figure A.4r - Extrinsic plot of $C_{bg}$ for a 100 μm by 100 μm (W/L) device, normalized by $C_{oe}$.
Figure A.4s - Intrinsic plot of $C_{bb}$ for a 100 µm by 100 µm (W/L) device, normalized by $C_{oi}$.

Figure A.4t - Extrinsic plot of $C_{bb}$ for a 100 µm by 100 µm (W/L) device, normalized by $C_{oe}$. 

- 259 -
Figure A.5a - Drain current versus $V_{GS}$ with $V_{BS}$ as a parameter for a 20 μm by 20 μm (W/L) device.

Figure A.5b - Drain current versus $V_{DS}$ with $V_{GS}$ as a parameter for a 20 μm by 20 μm (W/L) device.
Figure A.5c - Drain current versus $V_{GS}$ with $V_{DS}$ as a parameter for a 20 μm by 20 μm (W/L) device.

Figure A.5d - Transconductance versus $V_{GS}$ with $V_{DS}$ as a parameter for a 20 μm by 20 μm (W/L) device.
Figure A.5e – Intrinsic plot of $C_{gs}$ for a 20 µm by 20 µm (W/L) device, normalized by $C_{oi}$.

Figure A.5f – Extrinsic plot of $C_{gs}$ for a 20 µm by 20 µm (W/L) device, normalized by $C_{oe}$.

- 262 -
Figure A.5g - Intrinsic plot of $C_{gd}$ for a 20 μm by 20 μm (W/L) device, normalized by $C_{oi}$.

Figure A.5h - Extrinsic plot of $C_{gd}$ for a 20 μm by 20 μm (W/L) device, normalized by $C_{oe}$.
Figure A.5i - Intrinsic plot of $C_{gb}$ for a 20 $\mu$m by 20 $\mu$m (W/L) device, normalized by $C_{oi}$.

Figure A.5j - Extrinsic plot of $C_{gb}$ for a 20 $\mu$m by 20 $\mu$m (W/L) device, normalized by $C_{oe}$.
Figure A.5k – Intrinsic plot of $C_{gg}$ for a 20 µm by 20 µm (W/L) device, normalized by $C_{oi}$.

Figure A.5l – Extrinsic plot of $C_{gg}$ for a 20 µm by 20 µm (W/L) device, normalized by $C_{oe}$. 

- 265 -
Figure A.5m - Intrinsic plot of $C_{bs}$ for a 20 µm by 20 µm (W/L) device, normalized by $C_{oi}$.

Figure A.5n - Extrinsic plot of $C_{bs}$ for a 20 µm by 20 µm (W/L) device, normalized by $C_{oe}$.
Figure A.5o - Intrinsic plot of $C_{bd}$ for a 20 μm by 20 μm (W/L) device, normalized by $C_{oi}$.

Figure A.5p - Extrinsic plot of $C_{bd}$ for a 20 μm by 20 μm (W/L) device, normalized by $C_{oe}$.
Figure A.5q - Intrinsic plot of $C_{bg}$ for a 20 μm by 20 μm (W/L) device, normalized by $C_{oi}$.

Figure A.5r - Extrinsic plot of $C_{bg}$ for a 20 μm by 20 μm (W/L) device, normalized by $C_{oe}$.
Figure A.5s - Intrinsic plot of $C_{bb}$ for a 20 μm by 20 μm (W/L) device, normalized by $C_{oi}$.

Figure A.5t - Extrinsic plot of $C_{bb}$ for a 20 μm by 20 μm (W/L) device, normalized by $C_{oe}$. 

- 269 -
Figure A.6a - Drain current versus $V_{GS}$ with $V_{BS}$ as a parameter for a 20 μm by 4 μm (W/L) device.

Figure A.6b - Drain current versus $V_{DS}$ with $V_{GS}$ as a parameter for a 20 μm by 4 μm (W/L) device.
Figure A.6c - Drain current versus \( V_{GS} \) with \( V_{DS} \) as a parameter for a 20 \( \mu \)m by 4 \( \mu \)m (W/L) device.

Figure A.6d - Transconductance versus \( V_{GS} \) with \( V_{DS} \) as a parameter for a 20 \( \mu \)m by 4 \( \mu \)m (W/L) device.
Figure A.6e - Intrinsic plot of $C_{gs}$ for a 20 $\mu$m by 4 $\mu$m (W/L) device, normalized by $C_{oi}$.

Figure A.6f - Extrinsic plot of $C_{gs}$ for a 20 $\mu$m by 4 $\mu$m (W/L) device, normalized by $C_{oe}$. 
Figure A.6g - Intrinsic plot of $C_{gd}$ for a 20 μm by 4 μm (W/L) device, normalized by $C_{oi}$.

Figure A.6h - Extrinsic plot of $C_{gd}$ for a 20 μm by 4 μm (W/L) device, normalized by $C_{oe}$.
Figure A.6i - Intrinsic plot of $C_{gb}$ for a 20 $\mu$m by 4 $\mu$m (W/L) device, normalized by $C_{oi}$.

Figure A.6j - Extrinsic plot of $C_{gb}$ for a 20 $\mu$m by 4 $\mu$m (W/L) device, normalized by $C_{oe}$.
Figure A.6k - Intrinsic plot of $C_{gg}$ for a 20 μm by 4 μm (W/L) device, normalized by $C_{oi}$.

Figure A.6l - Extrinsic plot of $C_{gg}$ for a 20 μm by 4 μm (W/L) device, normalized by $C_{oe}$.
Figure A.6m - Intrinsic plot of $C_{bs}$ for a 20 μm by 4 μm (W/L) device, normalized by $C_{oi}$.

Figure A.6n - Extrinsic plot of $C_{bs}$ for a 20 μm by 4 μm (W/L) device, normalized by $C_{oe}$.
Figure A.60 - Intrinsic plot of $C_{bd}$ for a 20 μm by 4 μm (W/L) device, normalized by $C_{oi}$.

Figure A.6p - Extrinsic plot of $C_{bd}$ for a 20 μm by 4 μm (W/L) device, normalized by $C_{oe}$. 
Figure A.6q - Intrinsic plot of $C_{bg}$ for a 20 $\mu$m by 4 $\mu$m (W/L) device, normalized by $C_{oi}$.

Figure A.6r - Extrinsic plot of $C_{bg}$ for a 20 $\mu$m by 4 $\mu$m (W/L) device, normalized by $C_{oe}$.
Figure A.6s - Intrinsic plot of $C_{bb}$ for a 20 μm by 4 μm (W/L) device, normalized by $C_{oi}$.

Figure A.6t - Extrinsic plot of $C_{bb}$ for a 20 μm by 4 μm (W/L) device, normalized by $C_{oe}$. 

- 279 -
Figure A.7a - Drain current versus $V_{GS}$ with $V_{BS}$ as a parameter for a 4 μm by 20 μm (W/L) device.

Figure A.7b - Drain current versus $V_{DS}$ with $V_{GS}$ as a parameter for a 4 μm by 20 μm (W/L) device.
Figure A.7c - Drain current versus $V_{GS}$ with $V_{DS}$ as a parameter for a 4 μm by 20 μm (W/L) device.

Figure A.7d - Transconductance versus $V_{GS}$ with $V_{DS}$ as a parameter for a 4 μm by 20 μm (W/L) device.
Figure A.7e - Intrinsic plot of $C_{gs}$ for a 4 $\mu$m by 20 $\mu$m (W/L) device, normalized by $C_{oi}$.

Figure A.7f - Extrinsic plot of $C_{gs}$ for a 4 $\mu$m by 20 $\mu$m (W/L) device, normalized by $C_{oe}$. 

- 282 -
Figure A.7g - Intrinsic plot of $C_{gd}$ for a 4 \( \mu \)m by 20 \( \mu \)m (W/L) device, normalized by $C_{oi}$.

Figure A.7h - Extrinsic plot of $C_{gd}$ for a 4 \( \mu \)m by 20 \( \mu \)m (W/L) device, normalized by $C_{oe}$. 
- 283 -
**Figure A.7i** - Intrinsic plot of $C_{gb}$ for a 4 μm by 20 μm (W/L) device, normalized by $C_{oi}$.

**Figure A.7j** - Extrinsic plot of $C_{gb}$ for a 4 μm by 20 μm (W/L) device, normalized by $C_{oe}$.
Figure A.7k - Intrinsic plot of $C_{gg}$ for a 4 $\mu$m by 20 $\mu$m (W/L) device, normalized by $C_{oi}$.

Figure A.71 - Extrinsic plot of $C_{gg}$ for a 4 $\mu$m by 20 $\mu$m (W/L) device, normalized by $C_{oe}$.
Figure A.7m - Intrinsic plot of $C_{bs}$ for a 4 µm by 20 µm (W/L) device, normalized by $C_{oi}$.

Figure A.7n - Extrinsic plot of $C_{bs}$ for a 4 µm by 20 µm (W/L) device, normalized by $C_{oe}$.
Figure A.7o - Intrinsic plot of $C_{bd}$ for a 4 µm by 20 µm (W/L) device, normalized by $C_{oi}$.

Figure A.7p - Extrinsic plot of $C_{bd}$ for a 4 µm by 20 µm (W/L) device, normalized by $C_{oe}$. 

- 287 -
Figure A.7q - Intrinsic plot of $C_{bg}$ for a 4 μm by 20 μm (W/L) device, normalized by $C_{oi}$.

Figure A.7r - Extrinsic plot of $C_{bg}$ for a 4 μm by 20 μm (W/L) device, normalized by $C_{oe}$.
Figure A.7s - Intrinsic plot of $C_{bb}$ for a 4 $\mu$m by 20 $\mu$m (W/L) device, normalized by $C_{oi}$.

Figure A.7t - Extrinsic plot of $C_{bb}$ for a 4 $\mu$m by 20 $\mu$m (W/L) device, normalized by $C_{oe}$.
Figure A.8a - Drain current versus $V_{GS}$ with $V_{BS}$ as a parameter for a 4 $\mu$m by 4 $\mu$m (W/L) device.

Figure A.8b - Drain current versus $V_{DS}$ with $V_{GS}$ as a parameter for a 4 $\mu$m by 4 $\mu$m (W/L) device.
Figure A.8c - Drain current versus $V_{GS}$ with $V_{DS}$ as a parameter for a 4 μm by 4 μm (W/L) device.

Figure A.8d - Transconductance versus $V_{GS}$ with $V_{DS}$ as a parameter for a 4 μm by 4 μm (W/L) device.
Figure A.8e - Intrinsic plot of $C_{gs}$ for a 4 µm by 4 µm (W/L) device, normalized by $C_{o1}$.

Figure A.8f - Extrinsic plot of $C_{gs}$ for a 4 µm by 4 µm (W/L) device, normalized by $C_{oe}$. 

- 292 -
Figure A.8g - Intrinsic plot of $C_{gd}$ for a 4 $\mu$m by 4 $\mu$m (W/L) device, normalized by $C_{oi}$.

Figure A.8h - Extrinsic plot of $C_{gd}$ for a 4 $\mu$m by 4 $\mu$m (W/L) device, normalized by $C_{oe}$.
Figure A.8i - Intrinsic plot of $C_{gb}$ for a 4 $\mu$m by 4 $\mu$m (W/L) device, normalized by $C_{oi}$.

Figure A.8j - Extrinsic plot of $C_{gb}$ for a 4 $\mu$m by 4 $\mu$m (W/L) device, normalized by $C_{oe}$.
Figure A.8k - Intrinsic plot of $C_{gg}$ for a 4 $\mu$m by 4 $\mu$m (W/L) device, normalized by $C_{oi}$.

Figure A.8l - Extrinsic plot of $C_{gg}$ for a 4 $\mu$m by 4 $\mu$m (W/L) device, normalized by $C_{oe}$.

- 295 -
REFERENCES


- 305 -


- 306 -