

Chapter 2

MOSFET: Basics, Characteristics, and Characterization

Samares Kar

Abstract This chapter attempts to provide a theoretical basis for the Metal Oxide (Insulator) Semiconductor (MOS/MIS) Structure and the MOS/MIS Field Effect Transistor (MOSFET/MISFET), their characteristics, and their characterization (parameter extraction); the theoretical treatment starts from the first principles. While deriving the mathematical relations, assumptions have been avoided as far as possible. A comprehensive treatment is included which covers the important aspects of the function, mechanism, and operation of the MOS/MIS devices; in particular topics have been covered which are relevant to all the later chapters of the book and which will aid in reading the rest of this book. We begin this chapter with the theory of the classical MOS structure (non-leaky and SiO_2 single gate dielectric) and the classical MOSFET and then graduate to the MOS structure and the MOSFET with the high-k gate stack and the high mobility channels. Various aspects of the MOS/MOSFET devices analyzed in this chapter include the energy band profiles, circuit representations, electrostatic analysis (charge–voltage and capacitance–voltage relations), drain current versus drain voltage relation, quantum-mechanical phenomena (wave function penetration, tunneling, carrier confinement), nature of the high-k gate stack traps, and the pseudo-Fermi function inside the gate stack and the occupancy of the gate stack traps. Features such as capacitance–voltage (C–V) characteristics, flat-band and threshold voltages (V_{FB} and V_{T}), V_{T} versus EOT characteristics permeate the chapters; hence these features and characteristics such as conductance–voltage (G–V) characteristics have been discussed. A significant part of this chapter contains topics which are rarely seen in the literature and are yet to be well understood. As these topics (composition of the high-k gate stack, nature of the high-k gate stack charges, effects of the degradation factors) are of vital significance for the progress of the high-k gate stack technology, we have tried to analyze these issues. The final part of this chapter treats the various methods available for characterization of the high-k gate

S. Kar (✉)

Department of Electrical Engineering, Indian Institute of Technology, Kanpur 208016, India
e-mail: skar@iitk.ac.in

stacks, in particular, for the determination of the trap parameters—trap density, trap energy, trap capture cross-section, and the trap location inside the gate stack.

2.1 Introduction

This chapter will attempt to cover the basics of the electrical properties and the electrical characteristics, in particular, the concepts and the theory, which will help in reading the different chapters of the book. Our approach will be the following:

1. We will define any topic (e.g. energy band diagrams), technical term (e.g. frequency dispersion) or parameter (e.g. transconductance) used in this chapter at the place of its introduction in textual and/or mathematical form.
2. We will outline the importance and/or usefulness and/or application of the technical entity.
3. Set up the boundary conditions and outline all the assumptions made for obtaining any closed-form (analytical) equation, beginning from the first principles; outline all the important steps and the logic inputs in the solution process, and if necessary cite the source of a more detailed mathematical treatment.

Our focus will be on simplicity, continuity, and a complete coverage, and on avoiding all unnecessary complexity in algebra. The basics of the Metal Oxide Semiconductor (MOS) structure and the MOS Field Effect Transistor (MOSFET) and their important electrical characteristics and parameters will be covered. Our treatment will focus on the high- k (k is the dielectric constant) gate stacks, and will also include the high mobility channels.

The material presented in this chapter is linked to and overlaps with almost all the following chapters of the book. The following is an indicative outline of the linkage:

1. Chapter 3: Quantum mechanical tunneling;
2. Chapter 4: Quantum mechanical tunneling;
3. Chapter 5: Flat-band voltage, interface traps, Metal Induced Gap States (MIGS), charge neutrality level, Capacitance–Voltage (C–V) characteristics;
4. Chapter 6: Flat-band voltage, interface traps, MIGS, charge neutrality level, C–V characteristics;
5. Chapter 7: Carrier mobility, drain current in the linear regime, C–V characteristics;
6. Chapter 8: Traps, trap time constant, oxygen vacancies;
7. Chapter 9: C–V characteristics;
8. Chapter 10: Quantum mechanical tunneling, C–V characteristics, flat-band voltage;
9. Chapter 12: C–V and Conductance–Voltage (G–V) characteristics, parameter extraction.

References [1–3] are quality text/reference books on MOS structures and MOSFET containing the SiO_2 single gate dielectric. Reference [4] is a classical paper on many important topics of the MOS structures, particularly, its conductance originating from the interface traps. These references together contain most of the details of analysis of the classical devices with the SiO_2 single gate dielectric.

2.2 MIS/MOS Structure: Single SiO_2 Gate Dielectric

The Metal-Insulator/Oxide-Semiconductor (MIS/MOS) structure is the most important component of the MOSFET/MISFET. Its basic function is to modulate (and change the conductivity type—p to n or n to p) the channel conductivity, thereby modulating the drain current, which flows by drift, provided a channel exists from the source to the drain. In the high-k transistors, the semiconductor is silicon or a higher mobility material, e.g. Ge, SiGe, GaAs, InGaAs, InAs, InAlAs, InP, GaN, InSb, HgTe, PbTe, and the insulator is multi-layered, whose core is a high-k material, e.g. HfO_2 , La_2O_3 , HfSiO , HfAlO , HfNO , HfSiON , ErTiO_5 , SrTiO_3 , LaScO_3 , LaAlO_3 , GdScO_3 , LaLuO_3 , $\text{La}_2\text{Hf}_2\text{O}_7$, Gd_2O_3 , La_2SiO_5 , SrHfO_3 . To increase the permittivity (the gate stack capacitance), and the channel mobility, the future trend is towards higher-k gate dielectrics, and higher-mobility substrates.

We begin our analysis of the MIS structure with the following assumptions:

1. The semiconductor is p-type silicon; i.e. the channel on this substrate will be an n-channel—NMOSFET.
2. The gate dielectric is a dry thermal SiO_2 , as is the case when the Equivalent Oxide Thickness (EOT) is more than 1.3 nm or so.
3. The leakage current through the gate dielectric is insignificant.

Later, we will analyze, what complications the following features of a gate stack introduce into the treatment:

1. A leaky gate dielectric or gate stack.
2. A gate stack with multiple dielectrics (both covalent and ionic).
3. A high mobility channel material.

The most important aids in understanding the electrical characteristics of the MOS structure and the MOSFET are their energy band diagrams and the circuit representations.

2.2.1 Metal-Semiconductor Contact (*Schottky Barrier*)

One may consider the metal-semiconductor (MS) contact (or junction) to be a subclass of the MOS/MIS structure where the oxide/insulator thickness is nil. This device is popularly known as a Schottky barrier or diode or contact, named after

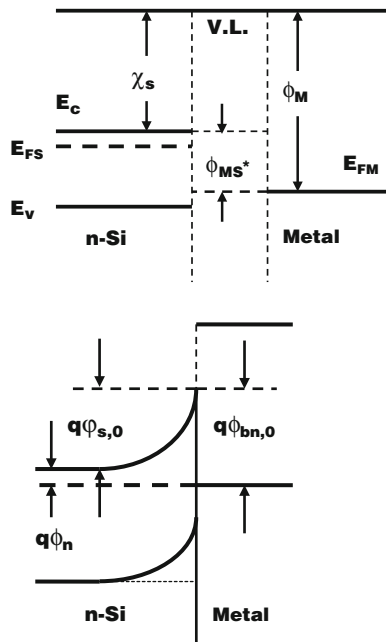
Walter Hermann Schottky, who made important contributions to the development of its theory, in particular the image force barrier lowering. Reference [5] is a good source for details on this topic. The Schottky diode itself is an important semiconductor device; in addition, it is an important component of the Schottky transistor (a bipolar junction transistor with a Schottky barrier between the base and the collector), Metal Semiconductor FET (MESFET in which the control electrode is the MS contact), High Electron Mobility Transistor (HEMT), and Schottky barrier carbon nano-tube (CNT) transistor. The MESFET and HEMT [6] have been the substitutes for the MOSFET/MISFET configuration on compound semiconductors, as earlier and still today it is difficult to realize high quality MOSFET/MISFET configurations on compound semiconductor substrates.

We have a more direct cause for looking at and analyzing the formation of the Schottky barrier and its basic nature. The MS contact is a limiting case of the MOS structure. The ultrathin MOS gate stack, with EOT less than 1.0 nm and scaling down to an EOT of 0.5 nm, is much closer to the MS contact than it is to the classical MOS structure. Therefore, an understanding of the MS contact is useful in analyzing the ultrathin MOSFET gate stack.

Many MS contacts may have an interfacial layer (say up to a nm thick) between the semiconductor and the metal [5]; however, it may be possible to avoid the interfacial layer as in a silicon/silicide MS contact [2]. When an intimate contact is established between the semiconductor and the metal, electron or hole exchange must take place between these two layers to equalize the energy and the Fermi level (What we call the Fermi level in semiconductor physics, is actually the chemical potential. At times, Fermi level is confused with the Fermi energy. The Fermi energy is defined only at absolute zero—0 K; it is the energy of the highest energy electron at 0 K). Figure 2.1a illustrates the energy band diagram across an MS contact before the intimate contact and Fig. 2.1b after the intimate contact. In the situation represented by Fig. 2.1a, the higher energy electrons in the conduction band are transferred from the n-type semiconductor to the metal, thereby setting up a dipole layer across the MS interface with the positive layer of ionized donors in the semiconductor sub-surface and the negative layer of transferred electrons on the metal surface.

The dipole layer sets up an electric field and a surface potential $\phi_{s,0}$ in the semiconductor space charge layer. The charge on the metal side must remain on its surface, as no electric field can penetrate a conductor. The metal surface charge density Q_M has to be equal and opposite of the semiconductor space charge density Q_{sc} , if the interface trap charge density Q_{it} is nil. The penetration of the electric field into a semiconductor depends upon its free carrier density and the charge on the semiconductor side resides in its space charge layer, a detailed analysis of which will be taken up later in Sect. 2.2.4. For a Schottky barrier, the semiconductor space charge is dominated by the charge of the ionized dopants. Perhaps the most important MS contact parameter is the Schottky barrier height, which dictates almost all the other parameters. The Schottky barrier height is the potential energy barrier perceived by a metal electron when transported from the metal to the semiconductor. The barrier height depends critically on the nature of the MS

Fig. 2.1 Energy band diagram across an n-semiconductor/metal MS contact (*top*) before and (*bottom*) after an intimate contact. E_c , E_v are the conduction, valence band edges, E_{FS} , E_{FM} are the semiconductor, metal Fermi levels, q is the electron charge, ϕ_M is the metal work function, χ_s is the semiconductor electron affinity, $\phi_{bn,0}$ is the zero-bias Schottky barrier height on n-type semiconductor, and $\phi_{s,0}$ is the zero-bias surface potential. V.L. is the vacuum level



interface—the interface trap density D_{it} . Two limits are invoked while discussing the Schottky barrier height—the Schottky–Mott limit in which case $D_{it} = 0$ and the Bardeen-limit in which case the high interface trap density pins the interface Fermi level. Let us take up the classical or the ideal case (Schottky–Mott limit) first. In this case, as represented by Fig. 2.1b, the barrier height for zero applied bias $\phi_{b,0}$ is given by:

$$\phi_{bn,0} = \frac{\phi_M - \chi_s}{q} = \phi_{s,0} + \phi_n \quad (2.1)$$

ϕ_n is the Fermi potential in n-semiconductor. Generally an interface state represents an allowed state inside the band-gap at the interface. A band-gap normally is an energy interval where the density of states is zero, i.e. where any states are forbidden. An interface represents a severe discontinuity in the periodic lattice; hence the solutions of the Schrödinger equation for the bulk crystal (with a periodic atomic arrangement and a periodic potential energy) do not obtain at the interface. Therefore allowed states may exist at the interface in the energy range which is a band-gap for the bulk crystal; in fact the entire distribution of the states at the interface may be different in the other energy ranges also, from what obtains in the bulk of the crystal. In addition to the intrinsic interface states, which are present due to the lattice mismatch at the interface, extrinsic states may also be present due to lattice defects and alien atoms, for which the interface is a center of attraction and a sink. In spite of a large amount of research carried out, the theoretical basis of the interface states is still incomplete [7].

Several concepts have been introduced in this area, one of which is the Metal Induced Gap States (MIGS), and one which we will come across in several chapters of the book. MIGS, which are one type of the intrinsic interface states, are due to the metal wave functions penetrating and decaying inside the semiconductor. A question related to the MIGS is whether the metal and the semiconductor wave functions can mix at the interface region. Charge Neutrality Level (CNL) is another concept which will be encountered often in connection with the interface traps. The interface trap charge density is zero or the interface trap charge is neutral if the Fermi level is at the CNL. Another related concept is the pinning parameter S , defined as $= \partial\phi_b / \partial\phi_M$. The Schottky-Mott limit corresponds to $S = 1$, i.e. $D_{it} = 0$, whereas the Bardeen limit corresponds to $S = 0$, i.e. $D_{it} = \infty$. When the interface trap density is infinite, then the interface traps pin the interface Fermi level so strongly that the Schottky barrier height becomes invariant of the metal work function, and in such a case, the barrier height is given by:

$$\phi_{bn,0} = \phi_M - \phi_{CNL} \quad (2.2)$$

ϕ_{CNL} is the CNL measured from the vacuum level. In most cases, the pinning parameter will lie between 0 and 1; in such cases, the barrier height will be given by:

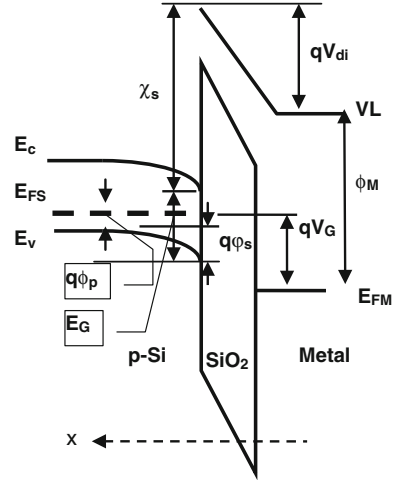
$$\phi_{bn,0} = (\phi_M - \phi_{CNL}) + S(\phi_{CNL} - \chi_s) \quad (2.3)$$

The Schottky barrier is a rectifying contact, i.e. a large current flows in the forward (a positive voltage applied on the metal with respect to an n-type semiconductor) direction, and a much smaller current in the reverse direction. Several mechanisms can operate simultaneously across an MS contact to transport carriers. In the case of a device quality Si/Metal Schottky barrier, the dominant carrier transport mechanism is likely to be thermionic emission over the potential energy barrier.

2.2.2 Energy Band Diagram

The profile of the energy bands along the axis of the applied electric field is very helpful in understanding the basic mechanisms of any semiconductor device. In these diagrams, the x axis represents most often the direction of the applied electric field, and the y axis the electron energy (generally electron energy increasing in the positive y direction and hole energy increasing in the negative y direction). Figure 2.2 illustrates the energy band profile of a p-Si/SiO₂/Metal structure across the x-axis, along which the gate voltage, V_G , is applied, for the equilibrium condition, i.e. the drain voltage $V_D = 0$. E_G is the semiconductor band gap, ϕ_s is the semiconductor band-bending (surface potential), and V_{di} is the potential across the gate dielectric. If we sum the energy parameters between the vacuum level (VL)

Fig. 2.2 Energy band profile across a p-Si/SiO₂/metal structure across the x direction, along which the gate voltage has been applied. VL is vacuum level, i.e. the energy an electron would have in absolute vacuum



and E_{FS} on both surfaces of the gate dielectric and equate, the following relation results:

$$\chi_s + E_G - q\phi_s - q\phi_p = qV_{di} + \phi_M - qV_G$$

Rearranging the terms, we obtain a relation for the gate voltage in terms of the potentials across the semiconductor and the gate dielectric, ϕ_s and V_{di} , respectively, and the work-function difference between the metal and the semiconductor, ϕ_{MS} .

$$V_G = \phi_s + V_{di} - [(\chi_s + E_G - \phi_M)/q - \phi_p] \quad (2.4)$$

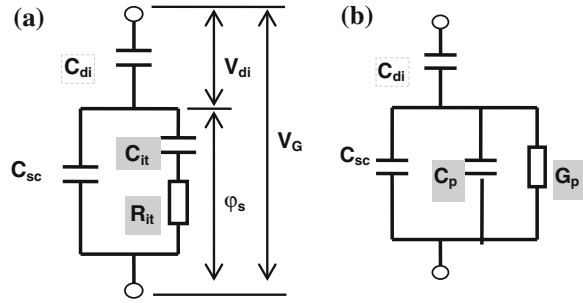
$$\phi_{MS} = (\chi_s + E_G - \phi_M)/q - \phi_p \quad (2.5)$$

In the ideal case, the work function difference is zero, and this is what one tries to achieve by a suitable choice of the metal work function and the semiconductor doping density.

2.2.3 Equivalent Circuit Representation

The salient parts of a device are generally reflected as elements in its circuit representation, which also has to be consistent with its energy band profile, as well as with the theoretical current voltage relations of the device. Figure 2.3 illustrates a relatively simple (assuming single-level interface states, equipotential semiconductor surface, etc.) equivalent circuit for the p-Si/SiO₂/Metal structure at an intermediate frequency. In Fig. 2.3, the gate dielectric is represented by the dielectric capacitance density, C_{di} , the semiconductor space charge layer is

Fig. 2.3 **a** Circuit representation of an MIS structure with a single gate dielectric, at an intermediate frequency, i.e. neither low nor high; **b** transformation of the circuit in **a**



represented by the semiconductor space-charge capacitance density, C_{sc} , and the interface traps by the interface trap capacitance density, C_{it} and the interface trap charging resistance, R_{it} , or alternatively by the equivalent parallel trap conductance density, G_p , and the equivalent trap capacitance density, C_p . The theoretical treatment is greatly simpler when the gate dielectric is a dry thermal SiO_2 layer, which for all practical purposes, is a near-perfect dielectric, i.e. is free of any bulk charges—both fixed (which do not charge or discharge in the operating voltage range) as well as trap charges. A perfect dielectric has only a dielectric capacitance inside, cf. (2.6), as in a plane parallel capacitor, and the electric field is constant inside, as reflected in Fig. 2.2 by the linear energy band profiles across the SiO_2 layer. Presence of bulk charges—both fixed and trap—create an additional non-linear potential and a varying electric field across the gate dielectric; additionally, the presence of bulk trap charges creates a trap capacitance in series with a trap resistance in parallel with the dielectric capacitance of the gate insulator. This issue will be further treated later, when we analyze the high- k gate stacks.

It may be noted that in Fig. 2.3a, the space charge layer in the semiconductor, which varies (charges or discharges) with the surface potential ϕ_s (i.e. the potential across this layer), is represented only by a capacitance (C_{sc}), while the interface traps (at the Si/SiO_2 interface) have been represented by a capacitance (C_{it}) in series with a charging resistor (R_{it}). In principle, any charging/discharging phenomenon is to be represented by a series RC branch, where C is the rate of charging/discharging ($\partial Q/\partial V$; Q is the charge and V is the potential.), R is the charging resistor, and $\tau (=RC)$ is the time constant. The series RC branch representation is the electrical engineering equivalent of the Kramers-Kronig relation in physics. In plain words, charging/discharging lags the application of a potential change by a typical time, called the relaxation time in physics and the time constant in electrical engineering. Often to simplify the analysis, we short-circuit the charging resistor, if the inverse of applied angular signal frequency, ω , is much larger than the relaxation time (i.e. the applied signal is easily followed by the charging/discharging process), and open-circuit it if the vice versa is the case. We have short-circuited the charging resistor for C_{sc} in Fig. 2.3, as the relaxation time (of the order of ps [3]) in this case is many orders of magnitude smaller than the inverse of the operating frequency (a few GHz).

The circuit representation in Fig. 2.3b results, when we transform the series $R_{it}C_{it}$ branch in Fig. 2.3a into a parallel G_p and C_p combination. We will see in later sections how the circuit of Fig. 2.3b is more appropriate for the analysis of the experimental interface state admittance data. The circuit representation of Fig. 2.3a simplifies to the one in Fig. 2.4a at high frequencies (f_h), and to the one in Fig. 2.4b at low frequencies (f_l). Any frequency much higher than the inverse of the inversion layer time constant τ_{inv} as well as the inverse of the minimum interface state time constant τ_{it} is a high frequency. Any frequency much smaller than the inverse of the inversion layer time constant τ_{inv} as well as the inverse of the maximum interface state time constant τ_{it} is a low (also called equilibrium) frequency.

The inversion layer time constant τ_{inv} and the interface state time constant τ_{it} are functions of the surface potential ϕ_s , and depend upon whether the device is an MOS capacitor or an MOSFET in operation. Both the low and the high frequency circuit representations of Fig. 2.4 are capacitive. At low frequencies, total MOS capacitance density C_{lf} is given by the relation:

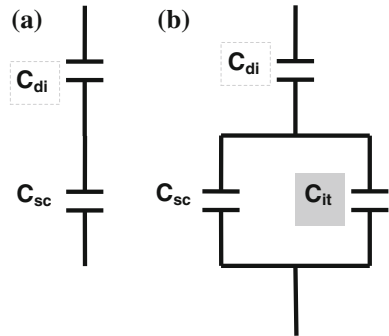
$$\frac{1}{C_{lf}} = \frac{1}{C_{di}} + \frac{1}{C_{sc} + C_{it}}; \text{ where } C_{di} = \frac{\epsilon_{di}}{t_{di}}; \text{ and } C_{it} = q \cdot D_{it} \quad (2.6)$$

ϵ_{di} is the dielectric permittivity, t_{di} is the dielectric thickness, and D_{it} is the interface state density. The simple relation for C_{it} in (2.6) follows from the definition: $C_{it} = \partial Q_{it} / \partial \phi_s$ (Q_{it} is the interface trap charge density.), as will be shown later. At high frequencies, the total MOS capacitance density C_{hf} is given by the relation:

$$\frac{1}{C_{hf}} = \frac{1}{C_{di}} + \frac{1}{C_{sc,hf}} \quad (2.7)$$

The relation for the high frequency space charge capacitance density $C_{sc,hf}$ depends upon whether the device is an MOS capacitor or an MOSFET. For an MOS capacitor, $C_{sc,hf}$ will reduce to the depletion space charge capacitance density C_{dep} , whereas for the MOSFET $C_{sc,hf}$ will be the same as C_{sc} . This point will be elaborated later.

Fig. 2.4 Simplification of the equivalent circuit in Fig. 2.3a at (a) high frequencies; and at (b) low frequencies or equilibrium



2.2.4 Electrostatic Analysis

The response of an equivalent circuit to an applied signal can be expressed once the mathematical relations for the different elements of the circuit are known which we now proceed to develop from an electrostatic analysis.

The electrostatic relations between potentials, charges, capacitances, and resistances, are much simpler, when the gate dielectric is a single near-perfect insulator like SiO₂. To expand the relation in (2.4), we need to substitute the relations for V_{di} and ϕ_p . The latter has the well-known relation:

$$\phi_p = kT/q \ln N_v/N_A = V_T \ln N_v/N_A \quad (2.8)$$

T is the absolute temperature, N_v is the effective density of states in the semiconductor valence band, N_A is the acceptor density, and V_T is thermal voltage. For a near-perfect gate dielectric, i.e. SiO₂, the expression for V_{di} is much simpler, while for a high- k gate stack (with as many as two bulk layers and three interfacial layers, full of fixed charges and traps), it can be very complicated. For the dry thermal SiO₂ gate dielectric, which has, for all practical purposes, no traps or fixed charges inside, V_{di} can be given by:

$$V_{di} = Q_M/C_{di} = -\frac{Q_{sc} + Q_{it} + Q_F}{C_{di}} \quad (2.9)$$

Q_F is the fixed charge density in the vicinity of the interface. It may be noted that Q_M on the metal surface is balanced by the charges on the silicon surface: $Q_{sc} + Q_{it} + Q_F$ and therefore has the opposite sign. The interface trap charge density can be expressed as:

$$Q_{it} = q \int (D_{itFD}^{Dch} - D_{itFD}^{Afe}) dE \quad (2.10)$$

D_{it}^D and D_{it}^A are respectively the density of the donor-type/acceptor-type interface states, f_{FD}^h and f_{FD}^e are respectively the hole/electron Fermi occupancy (Fermi-Dirac distribution), and E is energy. It can be easily shown that if the expression for Q_{it} is differentiated with respect to ϕ_s , the expression for C_{it} in (2.6) will result, since $D_{it} = D_{it}^D + D_{it}^A$.

Unfortunately, even in the case of the non-leaky Si/SiO₂/Metal system, an expression for Q_{sc} and C_{sc} (the remaining elements of the circuit representation) cannot be found in closed form, because of two main reasons:

1. In accumulation and strong inversion, one cannot use the Boltzmann distribution, but, needs to use the Fermi occupancy.
2. Carrier confinement (rather restrictions) in the x -direction (perpendicular to the interface) in the accumulation or the strong inversion potential well leads to standing waves (not Bloch functions, which are traveling waves), resulting in energy sub-bands inside the valence and conduction bands. This phenomenon

requires simultaneous application of the Poisson equation and the Schrödinger equation to carry out the electrostatic analysis.

In the case of a high- k gate stack, further complications arise:

- (a) Significant penetration of the semiconductor electron/hole wave-function occurs into the high- k gate stack and also transmission through the high- k gate stack to the metal takes place resulting in a high tunneling current, particularly for EOT in the 0.5–1.0 nm range.
- (b) Intermixing and interference between the semiconductor and the metal electron/hole wave-functions is possible.

The above complications notwithstanding, a closed-form solution promotes much clearer physical understanding. This is possible only if we ignore the carrier confinement effect as well as assume a Boltzmann distribution. The mathematical formulation and treatment in the following is generally credited to Ref. [8]. The starting point for the electrostatic analysis is the solution of the Poisson equation, for which we need an expression for the charge density per unit volume at a point x in the space charge layer, cf. Fig. 2.2, $\rho(x)$. The net charge density is the sum of the positive charge densities [ionized donor density, N_D^+ , and hole density, $p(x)$] minus the sum of the negative charge densities [ionized acceptor density, N_A^- , and electron density, $n(x)$]. We assume the doping density to be constant in the space charge layer.

$$\rho(x) = q \left[\sum N_D^+ - \sum N_A^- + p(x) - n(x) \right]$$

In the neutral regime, $\rho = 0$, which leads to the following simplification of the above relation:

$$\sum N_D^+ + p_0 = \sum N_A^- + n_0 \Rightarrow \rho(x) = q[(p - p_0) - (n - n_0)]$$

p_0/n_0 are the hole/electron concentrations in the neutral regime. The free carrier (electrons and holes) densities at a point x , $p(x)$, $n(x)$, will vary exponentially with the potential $\phi(x)$ at point x , and therefore across the space charge layer, cf. Fig. 2.5.

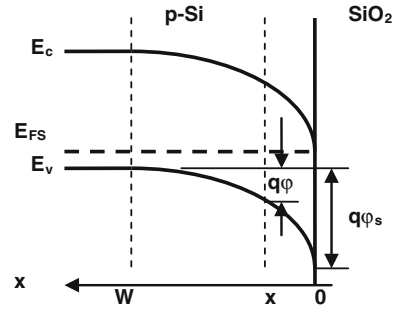
$$p(x) = p_0 \exp\{-\beta\phi(x)\}; \quad n(x) = n_0 \exp\{\beta\phi(x)\};$$

$\beta (=q/kT)$ is the inverse thermal voltage. The following steps follow from the application of the Poisson equation to the semiconductor space charge layer under the simplifications made:

$$d^2\phi/dx^2 = 1/2 d/d\phi \left(d\phi/dx \right)^2 = -q/\epsilon_s [p_0(e^{-\beta\phi} - 1) - n_0(e^{\beta\phi} - 1)]$$

One integrates the Poisson equation (the left side with respect to $d\phi/dx$, and the right side with respect to ϕ , after a slight rearrangement of the terms) to obtain the square of the electric field, $E(x) = -d\phi/dx$, at point x in the space charge layer.

Fig. 2.5 Energy band profile across the space charge layer for a p-Si/SiO₂ system in depletion



$$-d\phi/dx = E(x) = \pm \left[2qp_0/\beta\epsilon_s \left\{ (e^{-\beta\phi} + \beta\phi - 1) + n_0/p_0(e^{\beta\phi} - \beta\phi - 1) \right\} \right]^{1/2}$$

Subsequently one obtains the space charge density $Q_{sc} = -\epsilon_s E_i$, where ϵ_s is the semiconductor permittivity, and E_i is the electric field at the interface (also the semiconductor surface, $E_i = E(x)$ at $x = 0$, cf. Fig. 2.5):

$$Q_{sc} = \int_0^W \rho(x) dx = -\epsilon_s E_i$$

W is the space charge width, cf. Fig. 2.5. The negative sign in the above relation results from the fact that for a positive Q_{sc} , E_i is in the negative x direction, cf. Fig. 2.5. Since $E_i = E(\phi = \phi_s)$, the closed-form expression for Q_{sc} takes the form:

$$Q_{sc} = \mp \left[\frac{2q\epsilon_s N_A}{\beta} \left\{ (e^{-\beta\phi_s} + \beta\phi_s - 1) + n_0/p_0(e^{\beta\phi_s} - \beta\phi_s - 1) \right\} \right]^{1/2} \quad (2.11)$$

Differentiation of the space charge density with respect to the surface potential yields the closed-form mathematical relation for the space charge capacitance C_{sc} .

$$C_{sc} = \left| \frac{dQ_{sc}}{d\phi_s} \right| = \left(\frac{q\epsilon_s N_A \beta}{2} \right)^{1/2} \frac{\left| 1 - e^{-\beta\phi_s} + n_0/p_0(e^{\beta\phi_s} - 1) \right|}{\left[(e^{-\beta\phi_s} + \beta\phi_s - 1) + n_0/p_0(e^{\beta\phi_s} - \beta\phi_s - 1) \right]^{1/2}} \quad (2.12)$$

Equations (2.11) and (2.12) are valid in depletion (the majority carrier density < doping density but > intrinsic carrier density n_i , i.e. for a p-type semiconductor, $N_A > p_s > n_i$) and weak inversion (the minority carrier density < doping density but > intrinsic carrier density n_i , i.e. for a p-type semiconductor, $N_A > n_s > n_i$), as in these regimes, there is no potential well formation at the semiconductor surface and the attendant carrier confinement, and the Boltzmann distribution is valid. Although (2.11) and (2.12) are not valid in accumulation (the majority carrier density > doping density, i.e. for a p-type semiconductor, $p_s > N_A$) and strong inversion (the minority carrier density > doping density, i.e.

for a p-type semiconductor, $n_s > N_A$), these still are useful in providing some insight into the physical picture (p_s , n_s are hole, electron densities at the semiconductor surface).

The issues of quasi-thermal equilibrium, Fermi-Dirac occupancy, validity of such thermal-equilibrium entities as the law of mass action, etc., are important in the operation of high-k MOSFET/MISFET and ultrathin leaky gate dielectrics, but a clear analysis of these questions may be elusive. It may be noted that while deriving (2.11) and (2.12), we have assumed the law of mass action to be valid in the x direction in the space charge layer, i.e. $p(x)n(x) = n_i^2$. This is a good assumption even in an MOSFET in operation, if the gate dielectric is non-leaky.

Equations (2.1)–(2.12), would enable determination of the gate voltage V_G and the total low frequency and the high frequency MOS capacitance densities, C_{lf} and C_{hf} , respectively, for any value of the surface potential φ_s in depletion and weak inversion regimes. Equations (2.11) and (2.12) can be significantly simplified in depletion and in weak inversion. For a p-type semiconductor both in depletion and in weak inversion, φ_s is > 0 ; hence $\exp(-\beta\varphi_s) \ll \beta\varphi_s \ll \exp(\beta\varphi_s)$, and $\beta\varphi_s \gg 1$. Moreover, $(n_0/p_0)\exp(\beta\varphi_s) = n_s/p_0 \ll 1$. Under these conditions, (2.11) and (2.12) approximate to:

$$Q_{sc} \approx -(2q\varepsilon_s N_A \varphi_s)^{1/2} \quad (2.13)$$

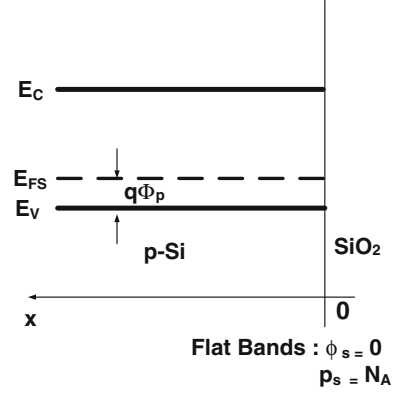
$$C_{sc} \approx \left(q\varepsilon_s N_A / 2\varphi_s \right)^{1/2} \quad (2.14)$$

Equations (2.13) and (2.14) are good approximations in both depletion and in weak inversion. It is worth noting that under these conditions, the semiconductor space charge layer reduces to a dielectric capacitor, whose capacitance density is given by ε_s/W , as there are only ionized dopants (whose charge is fixed) in this layer, and the free carrier density is insignificant; so, there is no charging or discharging, when φ_s is varied.

2.3 Flat-Band Voltage and Threshold Voltage: Single SiO₂ Gate Dielectric

Two very important MOS and MOSFET parameters are the flat-band voltage V_{FB} , corresponding to the condition: $\varphi_s = 0$ (or $p_s = N_A$ for a p-type semiconductor), and the threshold voltage V_T , corresponding to the onset of the strong inversion regime ($n_s = N_A$, for a p-type semiconductor). The flat-band and the onset of strong inversion conditions are illustrated by the energy profiles in Figs. 2.6 and 2.7, respectively. Equations (2.4), (2.5), and (2.9) yield the following expression for V_{FB} : which will be valid only for a gate dielectric free of bulk charges, cf. Fig. 2.6.

Fig. 2.6 Energy band profile across the space charge layer for a p-Si/SiO₂ system at flat band condition



$$V_{FB} = -[(Q_{it,fb} + Q_F)/C_{di}] - \phi_{MS} = -\left[(Q_{it,fb} + Q_F) \frac{t_{di}}{\epsilon_{di}}\right] - \phi_{MS} \quad (2.15)$$

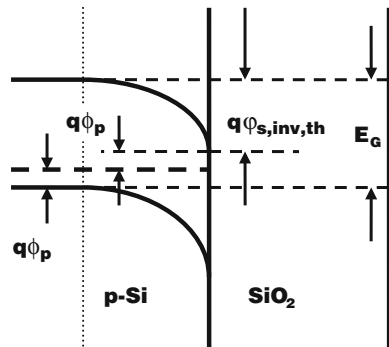
$Q_{it,fb}$ is the interface trap charge density at flat-band (Often, $Q_{it,fb}$ is ignored, and only Q_F is considered in calculating V_{FB}). Ideally, V_{FB} should be zero; in practice, one tries to obtain a value for V_{FB} as close as possible to zero. The flat-band voltage is for the MOSFET/MISFET a performance parameter, as it relates to the threshold voltage of the transistor, and also is a quality indicator, as it reflects the magnitude of the unwanted entities such as the trap density and the fixed charges. The device quality SiO₂ is a near-perfect gate dielectric; the experimental plot of the flat-band voltage V_{FB} versus the dielectric thickness t_{di} of an MOS structure with such a gate dielectric has been demonstrated to be a linear characteristic, as (2.15) indicates, yielding accurate values of the Si/metal work function difference and the interface charge density at flat-band, ($Q_{it,fb} + Q_F$) [9, 10].

Equations (2.4) and (2.9) yield the following expression for the threshold voltage (MOSFET turn-on voltage) V_T : cf. Figs. 2.7 and 2.2:

$$\begin{aligned} V_T &= \varphi_{s,inv,th} + V_{di,inv,th} - \phi_{MS} \\ &= \left(\frac{E_G}{q} - 2\phi_p\right) - \left(\frac{Q_{sc,inv,th} + Q_{it,inv,th} + Q_F}{C_{di}}\right) - \phi_{MS} \end{aligned} \quad (2.16)$$

$\varphi_{s,inv,th}$ is the surface potential, $V_{di,inv,th}$ is the potential across the gate stack, $Q_{sc,inv,th}$ is the semiconductor space charge density, and $Q_{it,inv,th}$ is the interface trap charge density, at the onset of strong inversion. At the onset of strong inversion, the Fermi level at the interface is $q\phi_p$ below the minority carrier band edge E_c , cf. Fig. 2.7. Hence the surface potential at the onset of strong inversion is given by: $\varphi_{s,inv,th} = (E_G/q) - 2\phi_p$.

Fig. 2.7 Energy band profile across the space charge layer for a p-Si/SiO₂ system at the onset of strong inversion



2.4 Capacitance–Voltage (C–V) Characteristics of the Si/SiO₂/Metal Structures

The capacitance–voltage, the C–V, characteristic of the MOS structure is perhaps the most frequently used tool in the characterization of both the MOS and the MOSFET devices; perhaps, the most important reasons are:

1. The ease with which the C–V characteristic can be measured, even in the case of high gate stack leakage current.
2. The accuracy with which it can be measured.
3. The sensitivity with which it reflects the defects, traps, and other deficiencies of the MOS structure and the MOSFET. In other words, the C–V curve is the most direct image of the MOS quality.
4. The large number of the important device parameters which can be extracted from this characteristic.

The fact that the C–V characteristic has been referred to throughout this book underscores the need to understand this characteristic accurately, which we will try to do in this section. Equations (2.4–2.12) enable us to understand a measured high/low frequency C–V characteristic. Such characteristics are displayed in Fig. 2.9 for a p-Si/SiO₂/Al capacitor, illustrated by the schematic of Fig. 2.8. The non-leaky, 5.2 nm thick SiO₂ layer was grown in dry O₂ at 1,100 °C. The high frequency characteristic was measured at 300 kHz, while the low frequency C–V was measured at a voltage ramp rate of 0.01 V/s. In Fig. 2.9, the accumulation, depletion, weak inversion, and the strong inversion regimes have been indicated. The surface potential needed to delineate these regimes was obtained from the Berglund integral [4] of the low frequency C–V curve. Figure 2.9 demonstrates one significant change as the gate dielectric becomes thinner, and the gate dielectric capacitance becomes comparable to the space charge capacitance in accumulation. This important change is the dominance of the accumulation and the strong inversion regimes over the depletion and the weak inversion regimes in the C–V characteristics.

Fig. 2.8 A schematic of a p-Si//SiO₂/Al MOS capacitor, whose low and high frequency C–V characteristics are displayed in Fig. 2.9

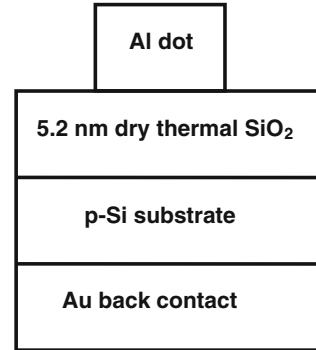


Fig. 2.9 Low and high frequency capacitance–voltage (C–V) characteristics of a p-Si/SiO₂/Al MOS capacitor. The dry thermal SiO₂ layer was 5.2 nm thick

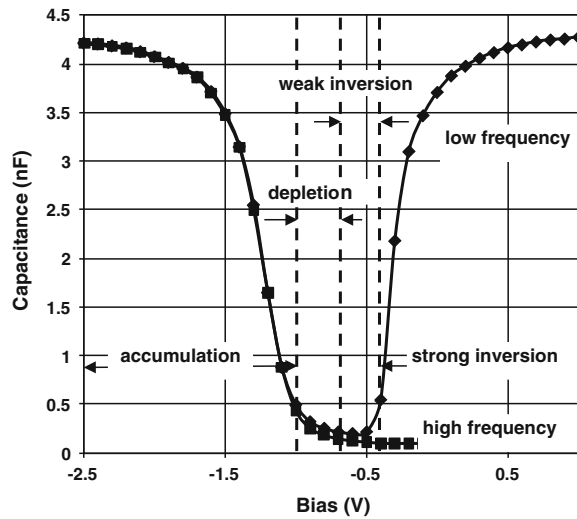


Figure 2.9 indicates that the capacitance is a weaker function of the bias in depletion and in weak inversion, and is a strong function of the bias in the early part of the accumulation and the early part of the strong inversion regimes; whereas the capacitance tends to saturate in the later part of accumulation and strong inversion. We will now proceed to qualitatively explain these features of the C–V characteristics. In this context, it may be noted that the C–V characteristics in Fig. 2.9 reflect a very low interface trap density, which means a small influence of the parameters C_{it} and Q_{it} on both the total C and the total V, cf. Figs. 2.3 and 2.4 and (2.11) and (2.12). Equations (2.13) and (2.14) show that in depletion and in weak inversion, both the space charge capacitance density C_{sc} and the space charge density Q_{sc} are parabolic (i.e. weak) functions of the surface potential. This is the main reason behind the slow variation of C with V in depletion and weak inversion.

For a p-type semiconductor in accumulation, φ_s is <0 ; hence $\exp(-\beta\varphi_s) \gg |\beta\varphi_s| \gg 1 \gg \exp(\beta\varphi_s)$. Under these conditions, (2.11) and (2.12) approximate to:

$$Q_{sc} \approx \sqrt{\frac{2q\epsilon_s N_A}{\beta}} \exp\left(-\frac{\beta\varphi_s}{2}\right); \quad C_{sc} \approx \sqrt{\frac{q\epsilon_s N_A \beta}{2}} \exp\left(-\frac{\beta\varphi_s}{2}\right) \quad (2.17)$$

Equation (2.17) indicates the space charge density, and therefore, the space charge capacitance density to be a strong exponential function of the surface potential in accumulation (It may be noted that $\beta\varphi_s$ is $\gg 1$). Even when the Fermi occupancy is used instead of the Boltzmann distribution and carrier confinement is considered, Q_{sc} and C_{sc} are still strong functions of the surface potential in accumulation. The mathematical relations for and the functional form of Q_{sc} and C_{sc} in accumulation and strong inversion will be discussed in a later section. The strong increase of C_{sc} as an accumulation layer grows is the reason why the capacitance C rises rapidly after accumulation sets in. As C_{sc} is in series with C_{di} , cf. Figs. 2.3 and 2.4, the total MOS capacitance tends to saturate to C_{di} , once C_{sc} becomes $\gg C_{di}$.

Once strong inversion sets in, for a p-type semiconductor, φ_s is >0 and also many times β^{-1} ; hence $\exp(\beta\varphi_s) \gg |\beta\varphi_s| \gg 1 \gg \exp(-\beta\varphi_s)$. Under these conditions, (2.11) and (2.12) approximate to:

$$Q_{sc} \approx -\sqrt{\frac{2q\epsilon_s n_0}{\beta}} \exp\left(\frac{\beta\varphi_s}{2}\right); \quad C_{sc} \approx \sqrt{\frac{q\epsilon_s n_0 \beta}{2}} \exp\left(\frac{\beta\varphi_s}{2}\right) \quad (2.18)$$

Equation (2.18) also indicates the space charge density, and therefore, the space charge capacitance density to be a strong exponential function of the surface potential in strong inversion (It may be noted that $\beta\varphi_s$ is $\gg 1$). Even when the Fermi occupancy is not approximated by the Boltzmann distribution and the carrier confinement is considered, Q_{sc} and C_{sc} are still strong functions of the surface potential in strong inversion. The strong increase of C_{sc} as a strong inversion layer grows is the reason why the capacitance C rises rapidly, once strong inversion sets in. As C_{sc} is in series with C_{di} , cf. Figs. 2.3 and 2.4, the total MOS capacitance tends to saturate to C_{di} , once C_{sc} becomes $\gg C_{di}$. The saturation of the MOS capacitance to the total gate dielectric capacitance C_{di} allows it to be extracted, and therefore also an EOT or Capacitive Equivalent Thickness (CET), from the measured C - V characteristic.

The MOS capacitance rises rapidly not only in strong inversion but also just before and at the onset of strong inversion, cf. Fig. 2.9. It may be noted that at the onset of strong inversion, for a p-type semiconductor, φ_s is >0 and also many times β^{-1} ; hence $\exp(\beta\varphi_s) \gg |\beta\varphi_s| \gg 1 \gg \exp(-\beta\varphi_s)$, further, we have $n_0 \exp(\beta\varphi_s)/p_0 = 1$, since $n_0 \exp(\beta\varphi_s) = n_s = p_0$. Under these conditions, (2.11) and (2.12) approximate to:

$$Q_{sc} \approx -\sqrt{2q\epsilon_s N_A \varphi_s}; \quad C_{sc} \approx \sqrt{\frac{2q\epsilon_s N_A}{\varphi_s}} \quad (2.19)$$

It is important to note that (2.19) yields a value for C_{sc} that is twice of what the depletion approximation, cf. (2.14), would yield. Depletion approximation ignores the contribution by the minority carriers. As the onset of strong inversion approaches, the minority carrier density approaches the doping density, and therefore, C_{sc} begins to increase. It may be noted that the MOS C–V has an asymmetry, and this asymmetry signifies whether the semiconductor is p- or n-type.

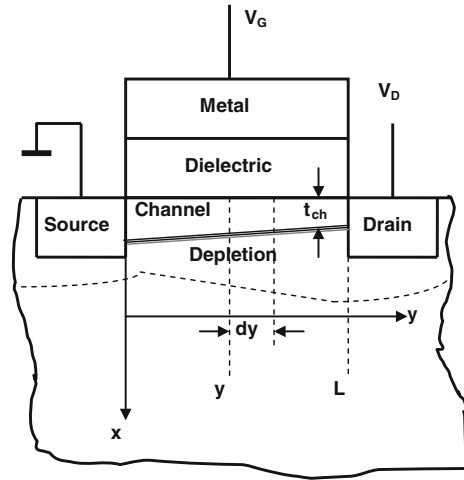
It is observed in Fig. 2.9, that the high frequency MOS capacitance saturates to a minimum capacitance in strong inversion. We now proceed to understand this phenomenon. At a high frequency, by the virtue of its definition, there is no contribution from the minority carriers to the space charge capacitance density C_{sc} . Once, a strong inversion layer forms and keeps growing, the surface potential ϕ_s changes very slowly with the bias V , since the space charge density Q_{sc} becomes very large, making the potential across the gate dielectric $\delta V_{di} \gg \delta \phi_s$, cf. (2.9). A near-constant ϕ_s makes C_{sc} saturate, cf. (2.14), which in series with C_{di} , renders a nearly constant minimum C in strong inversion. This minimum capacitance allows the doping density to be extracted from its measured value.

Perhaps, the most important parameter that the MOS C–V characteristics reflect is the interface trap density and its attendant effects. As Figs. 2.3 and 2.4 suggest, the difference between the low and the high frequency capacitance should directly yield the interface trap capacitance C_{it} , hence the interface trap density directly. Hence, the low frequency capacitance is taken to be a very useful and reliable indicator of the quality of the MOS capacitor and the MOSFET. The characteristics of Fig. 2.9 show no difference between C_{lf} and C_{hf} in accumulation and a only small difference in depletion and weak inversion, thereby demonstrating that this is indeed a high quality capacitor with a very low interface trap density. The mid-gap trap density obtained for this MOS structure from the MOS conductance data was about $2 \times 10^{10} \text{ cm}^{-2} \text{ V}^{-1}$.

2.5 Drain Current–Voltage Characteristics of MOSFET with SiO_2 Gate Dielectric

Our aim is to obtain a closed-form mathematical relation for the drain current I_D as a function of the drain voltage V_D with the gate voltage V_G as an important parameter, to gain an insight into the physical operation and behavior of the MOSFET. Application of the drain voltage along the channel, in the y direction, and of the gate voltage perpendicular to the channel and the semiconductor/insulator interface, in the x direction, cf. Fig. 2.10, requires a two-dimensional analysis, which cannot yield a closed-form solution.

Fig. 2.10 Idealized representation of the basic features of an MOSFET structure. The physical dimensions are disproportionate to allow emphasis on such elements of the MOSFET as the channel and the gate dielectric, both of which to a large extent dictate the MOSFET performance. Also, the contours of such elements as the channel and the depletion layer are only indicative. This representation assumes that the drain voltage $V_D < V_G$.



2.5.1 Ideal MOSFET: Linear Regime

Let us begin with a very simple one-dimensional analysis, which will provide some basic but important understanding of the device. Let us also characterize an ideal MOSFET gate stack with which we could later compare an MOSFET in reality and analyze the degradation of the channel parameters by the non-ideal factors. The following is a list of the non-ideal factors and we assume these to be absent in the ideal MOSFET gate stack:

1. Zero charges of any kind in the bulk and the interfaces of the gate insulator;
2. Zero semiconductor/metal work function difference;
3. Saturating inversion surface potential, i.e. the surface potential remains frozen at its value at the onset of the strong inversion regime and does no longer change with the gate voltage;
4. Low drain voltage compared to the gate voltage;
5. The charge of ionized dopants is zero.

In other words, in an ideal MOSFET, the entire gate voltage (100 % of it) is utilized in generating the channel charge Q_{ch} (consisting only of electrons in an n channel) and none of it is wasted on the non-ideal factors. The drain current is directly proportional to the channel charge Q_{ch} ; if parts of the gate voltage are wasted on the non-ideal factors and are therefore not available to generate Q_{ch} , then the drain current is reduced proportionately—this results in the degradation of the channel parameters by the non-ideal factors.

In an ideal MOSFET, under strong inversion, the source-channel-drain becomes a homogeneous semiconductor (with no junctions), allowing the majority carriers to flow by drift; hence the drain current is simply the drain voltage times the

channel conductance. For an n-channel MOSFET (i.e. p-type substrate), we could write, cf. Fig. 2.10:

$$\begin{aligned} I_D &= g_d V_D = \frac{W t_{ch}}{L} \sigma_{ch} V_D = \frac{W t_{ch}}{L} q n \mu_{ch} V_D = \frac{W}{L} |Q_{ch}| \mu_{ch} V_D \\ &= \frac{W}{L} C_{di} (V_G - V_T) \mu_{ch} V_D = \beta (V_G - V_T) V_D, \text{ where } \beta = \frac{W}{L} C_{di} \mu_{ch} \end{aligned} \quad (2.20)$$

g_d is the channel conductance, σ_{ch} is the channel conductivity, L is the channel length, cf. Fig. 2.10, W is the channel width, μ_{ch} is the channel mobility, Q_{ch} is the channel charge density, and β is the MOSFET quality factor. In deriving the above relation, we have assumed that the channel thickness t_{ch} is constant (due to low V_D) and is not a function of y or V_D , i.e. the channel conductance is a geometric one and is given by the channel area perpendicular to the y direction, $W t_{ch}$, the channel length L , and the channel conductivity σ_{ch} . We have assumed that the channel charge is dominated by the electrons; the channel charge density $Q_{ch} = q n t_{ch}$.

Since we had assumed zero bulk and interface charges for the gate dielectric, $Q_{it} = 0$ and $Q_F = 0$. Hence, $V_G = Q_M / C_{di} = -(Q_{sc} + Q_{it} + Q_F) / C_{di} = -Q_{sc} / C_{di}$. We assume that at the onset of strong inversion, i.e. when $V_G = V_T$, the space charge (i.e. basically the depletion charge) is negligible [assumption (5) above]. Therefore, any additional gate voltage, $(V_G - V_T)$ results in the creation of the channel charge, since the surface potential is frozen at its value at the onset of strong inversion, $\phi_{s,inv,th}$. Hence, $|Q_{ch}| = C_{di} (V_G - V_T)$. In strong inversion, the channel charge density Q_{ch} is predominantly that of the electrons. We must bear in mind that the relations in (2.20) are valid for very low drain voltages. The channel conductance g_d and the transconductance g_m can be expressed as:

$$g_D = \frac{\partial I_D}{\partial V_D} = \beta (V_G - V_T); \quad g_m = \frac{\partial I_D}{\partial V_G} = \beta V_D \quad (2.21)$$

There are some remarkable features of the relations in (2.20) and (2.21). These relations are very simple with a very simple derivation, still provide an insight into and understanding of the basic operation and performance of the MOSFET. The main significance of (2.20) could be summed as:

1. For low drain voltages (i.e. $V_D \ll V_G$), the $I_D(V_D)$ characteristic is linear. In other words, the ideal $I_D(V_D)$ relation is linear; any deviation from this linearity is a manifestation of the deviation from the ideal state and of degradation. We will see later how each of the five assumptions, made in deriving this relation, degrades and makes the drain current–voltage characteristic to droop and deviate from linearity.
2. Each of the five assumptions is an imperfection (a non-ideal factor) and causes the drain current to attenuate from its ideal value.
3. Ideally, the entire gate voltage, i.e. all of V_G , should be utilized for modulating the channel charge Q_{ch} . However, in a real MOSFET, each of the five non-ideal factors consumes a large part of the gate voltage.

4. Among the most important MOSFET performance parameters (Which directly influence the drain current, the switching speed, the channel conductance, and the transconductance.) are: (a) the channel mobility μ_{ch} ; and (b) the gate dielectric capacitance density C_{di} .
5. The slope ($\partial I_D / \partial V_D$) of the $I_D(V_D)$ characteristic can be used for the extraction of the important device parameters.

Equation (2.20) and the above remarks illustrate why in the case of the single gate dielectric (SiO_2 or SiNO), the technological trend was to reduce the gate dielectric thickness t_{di} , thereby increasing C_{di} . In the case of the high-k gate stack, the trend is both to increase the gate stack permittivity (by choosing insulators such as HfO_2 , La_2O_3 , thereby increasing C_{di} , and to enhance the channel mobility (by choosing semiconductors as Ge, GaAs, graphene).

2.5.2 Classical Model

We will now proceed to derive a more general and less ideal $I_D(V_D)$ relation, for which the first three assumptions of Sect. 2.5.1 will be retained, but the last two assumptions will be removed. The following mathematical treatment is generally credited to [1, 11, 12]. As soon as the drain voltage is applied, the uniformity of the channel disappears, i.e. the channel thickness t_{ch} becomes a function of the drain voltage, i.e. the channel can no longer be represented by a geometric conductance, but, becomes a V_D -dependent entity. This is what makes the $I_D(V_D)$ relation non-linear (As V_D increases in comparison to V_G , the $I_D(V_D)$ relation turns from linear to non-linear, and finally to saturation.), when V_D becomes significant. With the application of the drain voltage V_D , the voltage along the y direction, $V(y)$, increases from zero at the source ($y = 0$) to V_D at the drain ($y = L$), cf. Fig. 2.10. Consequently, the channel thickness $t_{ch}(y)$ is maximum at the source and minimum at the drain. At any point y, the voltage across the gate insulator is $[V_G - \varphi_s(y)]$. When $V_D = 0$, φ_s is not a function of y. But, for a non-zero V_D , φ_s is a function of y; such that at any point y, $\varphi_s(y) = \varphi_s(y = 0) + V(y)$, cf. Fig. 2.10.

It may be noted that for a non-zero drain voltage, the channel is in thermal non-equilibrium; consequently, the electron imref (i.e. the quasi-Fermi level) separates from the hole imref, as illustrated in Fig. 2.11. After a drain voltage has been applied, the voltage $V(y)$ equivalent energy, $qV(y)$, at a point y in the channel, must appear between the majority carrier imref in the neutral substrate and the majority carrier imref in the channel, as illustrated in Fig. 2.11. To maintain the channel and the strong inversion condition at any point y, for a p-Si substrate, the conduction band edge E_c has to maintain a certain minimum energy from the electron imref. Also, the surface potential at y for the strong inversion regime, $\varphi_{s,inv}(y)$, will exceed the surface potential at the source, $\varphi_{s,inv}(y = 0)$, by $V(y)$:

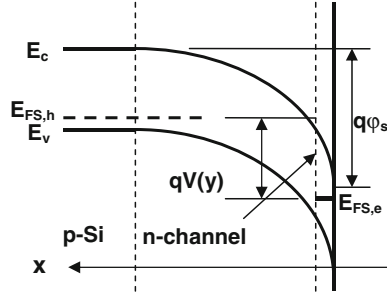


Fig. 2.11 Energy band profile of the semiconductor space charge region along the x -axis for the gate voltage $V_G > \text{threshold voltage } V_T$ (a channel exists at the semiconductor-dielectric interface), at the point y along the y direction (i.e. the direction of V_D) for a non-zero V_D . $E_{FS,h}$ is the hole (majority carrier) imref in the p-Si neutral region, while $E_{FS,e}$ is the electron imref in the channel. Note that electrons are minority carriers in the neutral p-Si region, but are majority carriers in the channel, i.e. after the p-Si substrate has been inverted

$\varphi_s(y) = \varphi_s(y = 0) + V(y)$. Therefore, the voltage across the gate dielectric $V_{di} = [V_G - \varphi_s(y = 0) - V(y)] = -Q_{sc}/C_{di}$. In other words, the voltage across the gate dielectric is maximum at the source and minimum at the drain, leading to a maximum channel charge and channel width at the source and a minimum channel charge and channel width at the drain. This makes the channel a non-geometric conductor, requiring a formulation of the current-voltage relation, different from the simple approach of Sect. 2.5.1.

As the channel thickness t_{ch} is now a function of y , we express the channel conductance Δg_d of an infinitesimal channel length dy , cf. Fig. 2.10, in the following manner:

$$\Delta g_d = \left(\frac{W}{dy} \right) \int_0^{t_{ch}} \sigma(x) dx, \quad \text{where } \sigma(x) = qn(x)\mu_e(x)$$

It may be noted that the electron density in an n-channel is a function of x , because of the potential $\varphi(x)$; hence the channel conductivity is also a function of x . The above relation may be rearranged as:

$$\Delta g_d = \left(\frac{W}{dy} \right) \mu_{ch} \int_0^{t_{ch}} qn(x) dx = \left(\frac{W}{dy} \right) \mu_{ch} |Q_{ch}|, \quad \text{where } |Q_{ch}| = q \int_0^{t_{ch}} n(x) dx$$

It may be noted that, in principle, both the channel carrier density n as well as the channel carrier mobility μ are functions of both x and y . (As will be explained later, we assume an effective channel mobility, to simplify the treatment).

The voltage across the channel element of length dy , $dV(y)$, can be represented as;

$$dV(y) = \frac{I_D}{\Delta g_d} = \frac{I_D}{W\mu_{ch}} \frac{dy}{|Q_{ch}|}$$

The above relation can be rearranged as:

$$\int_0^{V_D} |Q_{ch}| dV(y) = \left(\frac{I_D}{W\mu_{ch}} \right) \int_0^L dy = I_D L / (W\mu_{ch})$$

Hence, the drain current I_D can be expressed as:

$$I_D = \frac{W}{L} \mu_{ch} \int_0^{V_D} |Q_{ch}| dV(y)$$

In [Sect. 2.5.1](#), the charge of the ionized dopants was ignored [i.e. assumption (5)]. We will now formulate the channel charge without neglecting the space charge due to the ionized dopants.

$$\text{Since } V_G = \varphi_s + V_{di} = \varphi_s - \frac{Q_{sc}}{C_{di}}, \text{ we have: } Q_{sc} = -C_{di}(V_G - \varphi_s)$$

The channel charge density Q_{ch} may be equated to the total space charge density minus the charge density of the ionized dopants, for which, we may use the notation Q_{dep} ; to express the latter, we may use the depletion approximation, cf. [Sect. 2.2.4](#) and (2.13).

$$Q_{ch} = Q_{sc} - Q_{dep} = -C_{di}(V_G - \varphi_s) + \sqrt{2q\epsilon_s N_A \varphi_s} \quad (2.22)$$

It may be noted that for a p-type semiconductor, the depletion charge density Q_{dep} is negative, being due to the ionized acceptors. As the value of φ_s will be minimum at the source and maximum at the drain, the values of Q_{dep} will be the same. If we use the notation $\varphi_s(y=0)$ for the surface potential at the source, and the relation $\varphi_s(y) = \varphi_s(y=0) + V(y)$, then the drain current may be expressed as:

$$I_D = \frac{W}{L} \mu_{ch} \int_0^{V_D} \left[\{C_{di}(V_G - \varphi_s(y=0) - V(y))\} - \sqrt{2q\epsilon_s N_A [\varphi_s(y=0) + V(y)]} \right] dV(y)$$

Upon integration, we obtain a closed-form relation for the drain current:

$$I_D = \frac{W}{L} \mu_{ch} C_{di} \left[\left(V_G - \varphi_s(y=0) - \frac{V_D}{2} \right) V_D - \frac{2}{3} \gamma \left[(V_D + \varphi_s(y=0))^{\frac{3}{2}} - (\varphi_s(y=0))^{\frac{3}{2}} \right] \right] \quad (2.23)$$

$$\gamma = \frac{\sqrt{2q\epsilon_s N_A}}{C_{di}} = \frac{Q_{dep}}{C_{di}\phi_s^{1/2}} \quad (2.24)$$

Since we have assumed (assumption 3 in Sect. 2.5.1) that the surface potential remains frozen at its value at the onset of strong inversion, $\phi_s(y=0) = \phi_{s,inv,th} = (E_G/q) - 2\phi_p$.

As illustrated below, the relation in (2.23) reduces to the one in (2.20), for small values of V_D , i.e. $V_D \ll V_G$. It may be noted that the relation for the threshold voltage in (2.16) simplifies to what is written below for the assumptions made in Sect. 2.5.1, namely: $\phi_{MS} = 0$; $Q_{it} = 0$; $Q_F = 0$.

$$\begin{aligned} I_D &\approx \beta \left[\left(V_G - \phi_{s,inv,th} - \frac{V_D}{2} \right) V_D - \frac{2}{3} \gamma \left(\phi_{s,inv,th}^{3/2} + \frac{3}{2} \frac{V_D}{\phi_{s,inv,th}} \phi_{s,inv,th}^{3/2} - \phi_{s,inv,th}^{3/2} \right) \right] \\ &\approx \beta (V_G - \phi_{s,inv,th} - \gamma \sqrt{\phi_{s,inv,th}}) V_D \approx \beta (V_G - V_T) V_D, \quad V_T = \phi_{s,inv,th} + \gamma \sqrt{\phi_{s,inv,th}} \end{aligned}$$

When the drain voltage V_D is no longer small compared to the gate voltage V_G , the third term within the first parenthesis of (2.23), namely $V_D/2$, gains weight; consequently, the drain current does not increase rapidly with V_D , and begins to droop. The bracket in (2.23), which is multiplied by γ , is net positive and is to a lesser extent also responsible for the drain current deviating from a linear increase with the drain voltage and finally tending to saturate. It may be noted that the derivation of (2.23) is based upon the assumption that the channel exists in the entire region between the source and the drain, cf. Fig. 2.10, and that the carriers traverse the channel by drift. As the drain voltage increases, a situation comes about, when the channel disappears at the drain ($y = L$). This condition is known as pinch-off, which occurs first at drain, and with V_D increasing further, the pinch-off point y_p moves towards the source, cf. Fig. 2.12. The pinch-off condition can be defined as: $n_s = N_A$ (for p-type semiconductor) or $Q_{ch} = 0$. The saturation drain voltage, V_{DS} , is the drain voltage for which the pinch-off point $y_p = L$, and the corresponding drain current is the saturation drain current, I_{DS} . An expression for V_{DS} can be obtained by using (2.22) and the condition $Q_{ch}(y = L) = 0$:

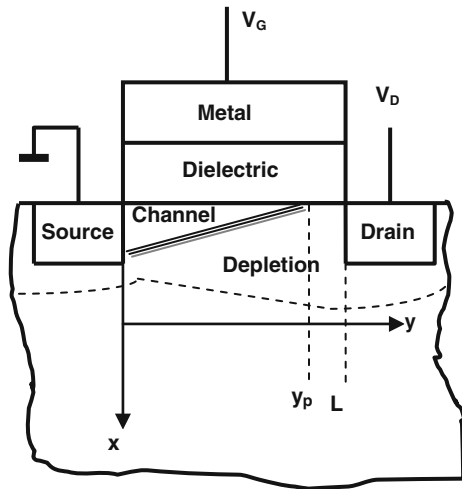
$$\begin{aligned} Q_{ch} &= -(V_G - V_{DS} - \phi_{s,inv,th}) C_{di} + \gamma C_{di} \sqrt{V_{DS} + \phi_{s,inv,th}} = 0 \\ \Rightarrow V_{DS} &= V_G - \phi_{s,inv,th} + \frac{\gamma^2}{2} \left[1 - \left(1 + \frac{4V_G}{\gamma^2} \right)^{\frac{1}{2}} \right] \end{aligned} \quad (2.25)$$

For ultrathin gate insulators, $\gamma \ll 1$, cf. (2.24). For example, for EOT = 1 nm and $N_A = 10^{17} \text{ cm}^{-3}$, $\gamma = 0.02 \sqrt{\text{V}}$. Equation (2.25) approximates, for $\gamma \ll 1$, to:

$$V_{DS} \approx V_G - V'_T, \quad \text{where } V'_T = \phi_{s,inv,th} + \gamma \sqrt{V_G} \quad (2.26)$$

Substitution of the relation for V_{DS} in (2.26) in (2.23), yields a relation for the saturation drain current I_{DS} versus the saturation drain voltage V_{DS} , for $\gamma \ll 1$:

Fig. 2.12 Schematic representation of an MOSFET, illustrating the onset of the saturation regime ($V_D > V_{DS}$), the pinch-off point y_p , and the attendant disappearance of the channel in the region between y_p and L



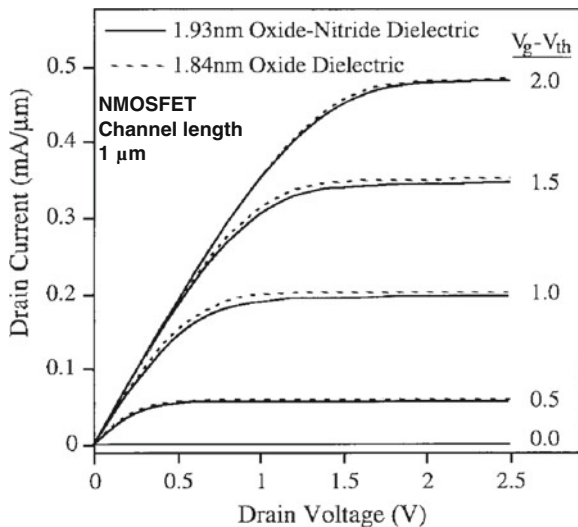
$$I_{DS} \approx \frac{\beta}{2} (V_G - V_T')^2 = \frac{\beta}{2} (V_{DS})^2 \quad (2.27)$$

In the saturation regime, the transconductance is given by:

$$g_m = \frac{\partial I_{DS}}{\partial V_G} = \beta (V_G - V_T') = \beta V_{DS} \quad (2.28)$$

Figure 2.13 presents the experimental drain current versus drain voltage characteristics of NMOSFETs: one with an ultrathin SiO_2 (EOT = 1.84 nm) and another with oxide-nitride (EOT = 1.93 nm) gate dielectric [13]. All the non-ideal

Fig. 2.13 Drain current versus drain voltage of n-channel MOSFET's with ultrathin oxide or oxide-nitride gate dielectric of EOT = 1.84 or 1.93 nm, respectively. The channel length was 1 μm . ($V_G - V_T$) varied between 0 and 2.0 V. Adapted from [13]



factors listed in Sect. 2.5.1 will be present in this MOSFET; however many of these factors will have much lower magnitudes than those in MOSFETs with high-k gate stacks. These measured characteristics of MOSFETs with silicon dioxide and silicon oxide-nitride dielectrics will be compared with those of MOSFETs with high-k gate stacks in Sect. 2.6.1.

2.6 High Dielectric Constant (k) Gate Stacks

In many ways (in physical, chemical, and electrical characteristics), the high permittivity gate dielectrics represent a drastic change, from the SiO₂ gate dielectric:

1. Dry thermal SiO₂ is a near-perfect dielectric, practically with no bulk charges and no space-charge capacitance; in that sense, high k materials are poor dielectrics with enormous charges ($>10^{13}/\text{cm}^2$); and high k gate stacks need to be represented by many (perhaps, as many as five) bulk trap and interface trap capacitances.
2. While the Si-SiO₂ interface is a true and marvelous gift of the nature (with an interface state density of the order of $10^{10}/\text{cm}^2/\text{V}$), the Si/high-k (and more so, Ge/high-k, GaAs/high-k, etc.) interface is a difficult one (with an interface state density $>10^{13}/\text{cm}^2/\text{V}$) with an uncertain chance of improvement.
3. With a poly-silicon gate electrode, the Si-SiO₂-poly-Si symmetric structure is practically immune to the kind of thermal and chemical stability problems encountered in the case of semiconductor/high-k-gate-stack/metal-electrode structures.
4. While the dry thermal SiO₂ owes its unmatched physical, chemical, and most importantly, electronic properties to its primarily covalent character, in sharp contrast to the high-k materials, which owe their many weaknesses to their primarily ionic character.
5. The covalent character of SiO₂ lends excellent matching with Si, while the ionic character of high-k is at the root of their poor matching with all semiconductors (including Si), which are all almost totally covalent (While Si and Ge are totally covalent, compound semiconductors are necessarily partly ionic).
6. SiO₂ is difficult to crystallize, and is a very stable material thermally. The opposite is true of the high-k materials, which crystallize easily much below the implant activation temperature.
7. Dry thermal silicon dioxide has the highest band-gap (of about 9 eV) among the inorganic solids, while the high-k materials have only moderate to high (say, 4–6 eV) band-gaps.
8. Dry thermal SiO₂ has the highest electrical resistivity (of the order of $10^{23} \Omega \text{ cm}$) ever recorded (The resistivity is purely electronic; there is no ionic contribution); the ionicity of the high-k materials lends them a conductivity, much higher than what their electronic or band-gap alone would suggest.

9. Dry thermal SiO₂ has perhaps the lowest dielectric constant (The electrical polarization is mainly electronic, with a very low ionic contribution.) among the inorganic oxides, whereas that of the high-k materials can be many times higher.

Succinctly expressed, everything would seem to be wonderful of the dry thermal SiO₂, except its abysmally low dielectric constant, whereas it is difficult to find a near-perfect property of the high-k gate dielectrics, except their high permittivity. It is hard to imagine how just one handicap undoes the dry thermal SiO₂, in spite of its truly amazing qualities, when gate dielectrics with sub-1-nm EOT are required. Fortunately, the high permittivity of the high-k gate stacks mitigates some of their problems:

1. Perhaps, the most effective among these relate to the huge reduction in the electric field across the different layers of the gate stack due to the high permittivity. It may be noted that a much lower electric field will induce the same inversion layer charge, as $Q_{ch} = \epsilon_{di}E_i$.
2. This, in turn, is responsible for the moderate potentials across the high-k gate stack layers, although, the high-k bulk and interface trap charge densities may be very large. Consequently, it has become possible to maintain the trend of reducing the threshold and the supply voltages.
3. The low electric fields may also promote gate stack reliability, even in the presence of huge charges inside.

We will discuss various aspects of the high-k gate stacks (physical structure and representation of the various layers, chemical nature, energy band profiles, nature and representation of the various bulk and interface traps, electrostatic analysis, circuit representations) in the later sections. First, we examine the implications of the high-k gate stack properties on the channel's electrical (drain current–voltage, etc.) characteristics.

2.6.1 Drain Current–Voltage Characteristics of MOSFETs with High-k Gate Stacks

We will now reexamine the validity of all the assumptions, made while deriving the classical $I_D(V_D)$ relation of (2.23), in the case of high-k gate stacks, and also examine how the drain current–voltage relation can be obtained in the closed form for MOSFETs with high-k gate stacks. The assumptions of Sect. 2.5.2 for the classical model were:

1. No semiconductor/metal work function difference, i.e. $\phi_{MS} = 0$.
2. No charges in the bulk of the gate stack.
3. No fixed oxide or interface state charges, i.e. $Q_{it} = 0$ and $Q_F = 0$.

4. Once strong inversion sets in, the inversion surface potential at the source remains frozen at the value $\varphi_{s,inv,th}$ and no longer increases with the gate voltage.

None of the above assumptions is tenable in the case of MOSFETs with high-k gate stacks, which are beset with enormous bulk charges and traps and work-function anomaly [14, 15]. After strong inversion sets in, the surface potential far from saturating has been observed to keep increasing continuously over the entire strong inversion regime [16, 17]. Hence the above assumptions taken together represent a serious contradiction to the reality that obtains in the case of the current high-k gate stacks.

Numerical models have been developed to make more realistic estimates of the drain current and the related parameters of MOSFETs. However, closed-form, text-book type equations, derived from the first principles, are needed to provide a sound physical insight into the critical parameters. Needed are mathematical relations in closed form for the drain current I_D , the channel conductance g_D , and the transconductance g_m which transparently illustrate how much the high-k gate stack charges, the non-saturating inversion surface potential, and the work function anomaly degrade the channel parameters. Even the recent semiconductor device and gate stack reference books still present the classical closed-form equations for the MOSFET channel parameters [18–20]. Our aim in this section will be four-fold, namely to: (1) have direct incorporation of the threshold-excess inversion surface potential ($\Delta\varphi_{s,inv} = \varphi_{s,inv} - \varphi_{s,inv,th}$), the total gate-stack charge density $Q_{di,gsc}$, and the semiconductor-metal work-function difference ϕ_{MS} in the equations for the drain current I_D , the channel conductance g_D , and the transconductance g_m ; (2) illustrate the scale of degradation of I_D , g_D , and g_m by the non-ideal factors of $\Delta\varphi_{s,inv}$, $Q_{di,gsc}$, and ϕ_{MS} , by quantitative analysis and estimation, using the available experimental data; (3) present text-book-appropriate relations in such a form that, how much each of the adverse factors degrades each of the channel parameters, becomes visible in a glance.

To derive the mathematical relations for the channel parameters, we will use the following general formulation for the drain current arrived at in Sect. 2.5.2:

$$I_D = \frac{W}{L} \mu_{ch} \int_0^{V_D} |Q_{ch}| dV(y) \quad (2.29)$$

The challenge is to find an expression for the channel charge $Q_{ch}(y)$ which would be valid in the case of the high-k gate stacks; in particular, the main problem is finding a realistic mathematical representation for the gate stack charge $Q_{di,gsc}(y)$, and the corresponding gate stack potential $V_{di}(y)$ and the channel charge $Q_{ch}(y)$, which will allow integration of (2.29) into a closed-form.

2.6.1.1 Gate Stack Potential $V_{di}(y)$

The gate stack potential V_{di} originates from the semiconductor space charge density Q_{sc} and all the charges in the gate stack. As Fig. 2.14 illustrates, the gate stack potential has many components (The energy band diagram of Fig. 2.14 will be discussed in detail in Sect. 2.6.3. It reflects experimental values of the surface potential and gate stack potentials and gate stack trap densities obtained from admittance-voltage-frequency and flat-band voltage versus EOT measurements over many MOS structures with varying EOT [17]. Figure 2.14 illustrates the complicated nature of the charge-potential relation inside a high-k gate stack). Unfortunately, at present, there is scarce information on the nature, location, and distribution of the charges in the high-k gate stacks [14–17]. Therefore, it is not possible to express realistically the potentials across the different layers and interfaces of the high-k gate stack, e.g. $V_{di,IL}$, $V_{di,dipole}$, in mathematical form as a function of y . We outline below an option for tackling this obstacle. We write down the total gate stack voltage as a sum of two components, one which is a strong function of y and the other which is relatively invariant of y .

$$V_{di}(y) = V_{di,sc}(y) + V_{di,gsc} \quad (2.30)$$

$V_{di,sc}$ is the total potential across the entire gate stack due to the semiconductor space charge density Q_{sc} , whereas $V_{di,gsc}$ is the total potential across the entire gate stack due to all the charges in the layers and the interfaces of the high-k gate stack. $V_{di,sc}(y)$ is a strong function of y , because of the strong variation in the surface potential $\phi_s(y)$ in the y -direction, cf. Sect. 2.5.2 and (2.31), causing the space charge density $Q_{sc}(y)$, which is a function of ϕ_s , to vary strongly.

$$\phi_{s,inv}(y) = \phi_{s,inv}(y=0) + V(y) = \phi_{s,inv,0,th} + \Delta\phi_{s,inv,0}(V_G) + V(y) \quad (2.31)$$

On the other hand, the bulk and the interface trap charges in the gate stack may not be a significant variant in the y -direction. The charge in traps inside the gate stack is decided by the occupancy of these traps; and the trap occupancy is decided by the pseudo-Fermi level inside the gate stack, cf. Fig. 2.14. The pseudo-Fermi level at a plane x inside the gate stack indicates the occupancy of traps at that plane and whether the traps at that plane is communicating more readily with the semiconductor surface, i.e. is controlled and given by $E_{FS,h}$, or more readily with the metal surface, i.e. is controlled and given by E_{FM} , or is communicating with neither, cf. Fig. 2.14 [12, 16]. To analyze their variation with y , the gate stack traps can be classified into three groups:

Group 1 The charge in traps at the metal surface and the charge in the traps inside the gate stack which exchange electrons more readily with the metal surface will not be a function of y , since this charge is controlled by the metal Fermi level E_{FM} , which is not a function of y , the potential on the metal surface being the same everywhere, cf. Fig. 2.14.

Group 2 There may be traps deep inside the gate stack which are unable to exchange electrons/holes either with the Si or the metal surface. Moreover, there

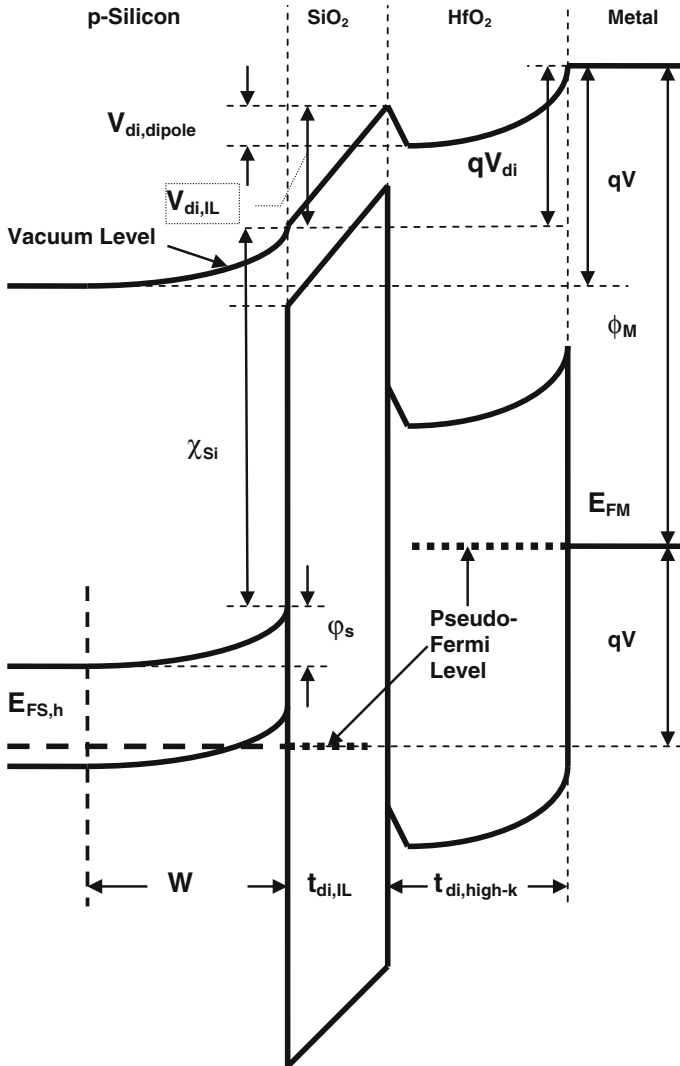


Fig. 2.14 A schematic representation of the energy profiles across a p-silicon/SiO₂/HfO₂/Ta₂N MOS capacitor in strong accumulation, under a bias of -1.82 V. The SiO₂ layer was about 1 nm and the HfO₂ layer about 2 nm thick, while the EOT was about 1.9 nm. $V_{di,dipole}$ is the potential drop across the dipole at the IL/high-k interface. $E_{FS,h}$ is the hole (majority carrier) imref. $t_{di,IL}$, $t_{di,high-k}$ are the thickness of the IL and the high-k layer, respectively. The pseudo-Fermi function is supposed to indicate the occupancy of the traps in the gate stack

may be traps in the gate stack whose energy levels remain much higher or much lower than the Si quasi-Fermi level $E_{FS,h}$, cf. Fig. 2.14. Effectively, these traps will act as fixed charges and will remain invariant of potential and therefore also of y .

Group 3 The charge in traps at the Si surface and the charge in traps inside the gate stack which exchange electrons more readily with the Si surface will be controlled by the Si quasi-Fermi level $E_{FS,h}$ the variation of this charge with y will depend upon the variation of $E_{FS,h}$ with y , cf. Figs. 2.11 and 2.14. The variation in the surface potential $\phi_{s,inv}$ will be $= V_D$ between the source and the drain according to (2.31). However, the variation of $E_{FS,h}$ with y will be a small fraction of eV_D , as for a channel (i.e. strong inversion) to exist, $E_{FS,h}$ has to remain close to E_c , cf. Fig. 2.11. Two other factors need to be considered to analyze how much the charge in group 3 traps will change with y . All experimental evidence consistently suggests that the trap density is lowest at the Si/IL interface, and the S/IL interface trap charge is a very small fraction of the total gate stack charges [14, 15, 17]. The magnitude of charge exchange between the Si surface and the traps inside the gate stack will be determined by the electron density at the Si surface, the wave function attenuation (depends on the conduction band offset, the tunneling electron mass, and the distance between the trap location and the Si surface), the trap capture/emission cross-section, and the frequency, cf. Fig. 2.14 [17], and may be a small fraction of the total gate stack charge.

Hence the charge of only the group 3 gate stack traps may vary with y and the error may be small if we ignore the variation with y of the charge in the group 3 traps, and consider $V_{di,gsc}$ to be invariant of y to enable its integration with respect to $V(y)$ into a closed-form expression. Combination of (2.4), (2.30), and (2.31) would then yield:

$$\begin{aligned} V_{di,sc}(y) &= V_G - \phi_{s,inv}(y=0) - V(y) - V_{di,gsc} - \phi_{MS,p} = -\frac{Q_{sc}(y)}{C_{di}} \\ &= -\frac{Q_{ch}(y) + Q_{dep}(y)}{C_{di}} \end{aligned} \quad (2.32)$$

or:

$$Q_{ch}(y) = -C_{di}[V_G - \phi_{s,inv}(y=0) - V(y) - V_{di,gsc} - \phi_{MS,p}] - Q_{dep}(y) \quad (2.33)$$

In (2.33), only $V(y)$ and $Q_{dep}(y)$ are functions of y . The work function difference (anomaly) may be considered invariant of y .

2.6.1.2 Drain Current Versus Drain Voltage Characteristic

Substitution of the expression for the channel charge density in (2.33) into the equation for the drain current in (2.29) yields:

$$\begin{aligned}
I_D &= \frac{W}{L} \mu_{ch} \int_0^{V_D} \left[\{C_{di} [V_G - \varphi_{s,inv,0} - V(y) - V_{di,gsc} - \phi_{MS,p}] \} + Q_{dep}(y) \right] dV(y) \\
&= \frac{W}{L} \mu_{ch} \int_0^{V_D} \left[\{C_{di} [V_G - \varphi_{s,inv,0} - V(y) - V_{di,gsc} - \phi_{MS,p}] \} - \sqrt{2q\epsilon_s N_A \varphi_s(y)} \right] dV(y) \\
&= \frac{W}{L} \mu_{ch} \int_0^{V_D} \left[\{C_{di} [V_G - \varphi_{s,inv,0} - V(y) - V_{di,gsc} - \phi_{MS,p}] \} - \sqrt{2q\epsilon_s N_A (\varphi_{s,inv,0} + V(y))} \right] dV(y)
\end{aligned} \tag{2.34}$$

$\varphi_{s,inv,0}$ is the inversion surface potential at $y = 0$, $\varphi_{s,inv}(y = 0)$. In (2.34), the classical depletion approximation relation has been used for Q_{dep} [3]. Even though all the non-ideal factors present in the high-k gate stack have been considered in (2.34), we have been able to represent these factors in such a manner that (2.34) can be integrated into a closed form to yield the following mathematical relation for the drain current:

$$I_D = \frac{W}{L} \mu_{ch} C_{di} \left[\left(V_G - \varphi_{s,inv,th,0} - \Delta\varphi_{s,inv,0} - V_{di,gsc} - \phi_{MS,p} - \frac{V_D}{2} \right) V_D - \frac{2}{3} \frac{\sqrt{2q\epsilon_s N_A}}{C_{di}} \left\{ (\varphi_{s,inv,th,0} + \Delta\varphi_{s,inv,0} + V_D)^{3/2} - (\varphi_{s,inv,th,0} + \Delta\varphi_{s,inv,0})^{3/2} \right\} \right] \tag{2.35}$$

The classical drain current versus the drain voltage relation, cf. (2.23) which was derived in Sect. 2.5.2 could be expressed as:

$$I_D = \frac{W}{L} \mu_{ch} C_{di} \left[\left(V_G - \varphi_{s,inv,th,0} - \frac{V_D}{2} \right) V_D - \frac{2}{3} \frac{\sqrt{2q\epsilon_s N_A}}{C_{di}} \left\{ (\varphi_{s,inv,th,0} + V_D)^{3/2} - (\varphi_{s,inv,th,0})^{3/2} \right\} \right] \tag{2.36}$$

Normalization of (2.35) by (2.36) yields a relation which directly illustrates the effects of the gate stack charges $V_{di,gsc}$, the non-saturating inversion surface potential $\Delta\varphi_{s,inv,0}$, and the work function difference $\phi_{MS,p}$ on the drain current:

$$\begin{aligned}
\frac{I_{D,high-k}}{I_{D,ideal}} &= \frac{\mu_{ch,high-k}}{\mu_{ch,ideal}} \times \left(V_G - \varphi_{s,inv,th,0} - \Delta\varphi_{s,inv,0} - V_{di,gsc} - \phi_{MS,p} - \frac{V_D}{2} \right) V_D - \\
&\quad \frac{\frac{2}{3} \gamma \left[(\varphi_{s,inv,th,0} + \Delta\varphi_{s,inv,0} + V_D)^{3/2} - (\varphi_{s,inv,th,0} + \Delta\varphi_{s,inv,0})^{3/2} \right]}{\left(V_G - \varphi_{s,inv,th,0} - \frac{V_D}{2} \right) V_D - \frac{2}{3} \gamma \left[(\varphi_{s,inv,th,0} + V_D)^{3/2} - (\varphi_{s,inv,th,0})^{3/2} \right]}
\end{aligned} \tag{2.37}$$

The following are the important features of the normalized drain current as represented by (2.37). As the channel conductance g_D and the transconductance g_m are directly linked to the drain current I_D , many of these features belong to the latter as well.

Non-Saturating Surface Potential—The surface potential does not saturate in the case of the ultrathin high-k gate stacks because of a number of reasons: (1) As the EOT is decreased, the gate stack dielectric capacitance density C_{di} increases and becomes comparable to the semiconductor space charge capacitance density

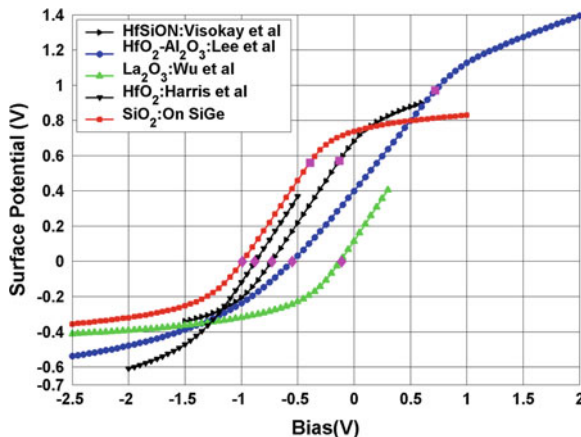


Fig. 2.15 Experimental surface potential versus applied gate bias plot extracted from the Berglund integral [60] of the measured equilibrium capacitance–voltage (C – V) characteristics of MOS structures on p-type silicon with five different gate stacks: HfSiON (*right faced triangle, black*), HfO₂-Al₂O₃ (*circle, blue*), La₂O₃ (*triangle, green*), HfO₂ (*inverted triangle, black*), and SiO₂ (*square, red*), and EOT values of 2.0, 1.7, 1.4, 0.5, and 3.9 nm respectively. Surface potential in strong inversion could be extracted only for three of the five MOS structures. The C – V data for four of the characteristics were taken from the literature: HfSiON [56], HfO₂-Al₂O₃ [57], La₂O₃ [58], HfO₂ [59]. It maybe noted that the *diamond marker* indicates the flat-band point, while the *square marker* indicates the onset of strong inversion

C_{sc} . Consequently, the surface potential change given by: $\Delta\varphi_{s,inv} = \Delta V_{di}C_{di}/C_{sc}$ becomes larger in comparison to the gate stack potential change ΔV_{di} . (2) Because of quantum-mechanical carrier confinement in the strong inversion layer, C_{sc} is significantly less than its classical value. (3) The very high trap density of high- k gate stacks makes ΔV_{di} larger. The degrading feature of the non-saturating surface potential has not been widely recognized in the literature. $\varphi_{s,inv,th,0}$ is the surface potential at the onset of strong inversion at the source: $\varphi_{s,inv,th,0} = E_G/q - 2\phi_p$ for a p-type semiconductor. In the classical formulation [1, 11, 12], the surface potential at the source was assumed to remain frozen at the value $\varphi_{s,inv,th,0}$ once the channel was formed. In the current high- k transistors, the inversion surface potential $\varphi_{s,inv,0}(V_G)$ has been observed to increase significantly with the gate voltage V_G ; experimental data indicate that the surface potential may increase by as much as 0.4 V after strong inversion has set in, i.e. $\Delta\varphi_{s,inv,0} = (\varphi_{s,inv,0} - \varphi_{s,inv,th,0}) = 0.4$ V [21, 22]. Figure 2.15 illustrates the strong variation of the surface potential in both accumulation and in strong inversion for four different high- k gate stacks in comparison with a control SiO₂ gate dielectric. The excess inversion surface potential $\Delta\varphi_{s,inv,0}$ appears twice in the numerator of (2.37), once inside the parenthesis, then within the bracket; consequently $\Delta\varphi_{s,inv,0}$ degrades (i.e. reduces) the drain current twice.

Gate Stack Charges and Traps—Flat-band voltage measurements and results from the conductance technique [17] indicate the charges and the traps in the current device-quality high-k gate stacks to be orders of magnitude higher (net gate stack charge density exceeding $q \times 10^{13} \text{ cm}^{-2}$) than what obtained in the case of the dry thermal SiO_2 gate dielectrics. Experimental data [17] indicate the high-k gate stack potential due to the gate stack charges $V_{\text{di,gsc}}(V_G)$ to be significant (hundreds of mV). $V_{\text{di,gsc}}(V_G)$ will reduce the drain current below its ideal value, cf. (2.37), and this reduction will increase if the gate stack charge increases, e.g. due to gate stack degradation [23, 24].

Work Function Difference Anomaly—The work-function difference term is absent in the classical drain current relation, cf. (2.36), (2.37). Unfortunately, the high-k gate stacks suffer from a serious work-function anomaly and lack of control [25, 26]; consequently the desired work function cannot be realized, and the work-function difference remains not only significant (a few hundred mV) but also changes with subsequent processing. A significant ϕ_{MS} will decrease the drain current below its ideal value and its lack of control introduces drain current instability, cf. (2.35), (2.37).

2.6.1.3 Estimate of Drain Current Degradation

All the three factors discussed above degrade the high-k gate stack drain current $I_{\text{D,high-k}}$ to a fraction of its ideal value, $I_{\text{D,ideal}}$, cf. (2.37). It may be noted that the values of the parameters V_G , $\varphi_{\text{s,inv,th,0}}$, V_D , and γ in (2.37) and (2.24) may be obtained from the device design, whereas the values of the parameters $\Delta\varphi_{\text{s,inv,0}} = (\varphi_{\text{s,inv,0}} - \varphi_{\text{s,inv,th,0}})$, $V_{\text{di,gsc}}$, and ϕ_{MS} may be extracted from carefully planned experiments on the MOS device. It may be instructive to make a rough estimate of the normalized drain current in (2.37), and obtain a feel for the importance of the various degrading factors. The body effect parameter γ is proportional to EOT and to $(N_A)^{1/2}$; choosing $\text{EOT} = 1.0 \text{ nm}$ and $N_A = 3.55 \times 10^{17} \text{ cm}^{-3}$ yields a round figure of 0.1 for γ . We may choose: a gate voltage $V_G = 2.2 \text{ V}$; a drain voltage $V_D = 0.4 \text{ V}$ and the surface potential at the source at the onset of strong inversion $\varphi_{\text{s,inv,th,0}} = 0.9 \text{ V}$. For the estimation, it may be assumed that the increase in the surface potential after the onset of strong inversion $\Delta\varphi_{\text{s,inv,0}} = (\varphi_{\text{s,inv,0}} - \varphi_{\text{s,inv,th,0}}) = 0.2 \text{ V}$ [21, 22]. Flat-band voltage versus EOT measurements indicate that the total gate stack potential due to all its fixed and trap charges may be as much as 0.3 V [17]; we may choose a gate stack potential due to the net gate stack charges $V_{\text{di,gsc}} = 0.2 \text{ V}$. The work function anomaly varies a great deal depending upon the metal, the gate stack high-k layer, and the processing. At the current state of the work function control, it may be reasonable to assume a work-function difference $\phi_{\text{MS,p}} = 0.1 \text{ V}$. Hence, a numerical estimate of the drain current according to (2.37) would be:

$$\begin{aligned}
\frac{I_{D,high-k}}{I_{D,ideal}} &= \frac{\mu_{ch,high-k}}{\mu_{ch,ideal}} \times \frac{(2.2 - 0.9 - 0.2 - 0.2 - 0.1 - 0.2)0.4 - \frac{2}{3} \cdot 0.1 \left[(0.9 + 0.2 + 0.4)^{\frac{3}{2}} - (0.9 + 0.2)^{\frac{3}{2}} \right]}{(2.2 - 0.9 - 0.2)0.4 - \frac{2}{3} \cdot 0.1 \left[(0.9 + 0.4)^{\frac{3}{2}} - (0.9)^{\frac{3}{2}} \right]} \\
&= \frac{\mu_{ch,high-k}}{\mu_{ch,ideal}} \times \frac{0.6 \cdot 0.4 - 0.067 \left((1.5)^{3/2} - (1.1)^{3/2} \right)}{1.1 \cdot 0.4 - 0.067 \left[(1.3)^{\frac{3}{2}} - (0.9)^{\frac{3}{2}} \right]} = \frac{\mu_{ch,high-k}}{\mu_{ch,ideal}} \times \frac{0.24 - 0.067 \cdot 0.69}{0.44 - 0.067 \cdot 0.63} \\
&= \frac{\mu_{ch,high-k}}{\mu_{ch,ideal}} \times \frac{0.24 - 0.04}{0.44 - 0.04} = \frac{0.20}{0.40} \frac{\mu_{ch,high-k}}{\mu_{ch,ideal}} = 0.50 \frac{\mu_{ch,high-k}}{\mu_{ch,ideal}}
\end{aligned} \tag{2.38}$$

These calculations suggest that the part of the relation containing the bracket in (2.37) or (2.38), is not significant for small drain voltages; the part containing the bracket represents the difference between the depletion charge at the drain and at the source and assumes importance for large drain voltages. Therefore, for the triode regime, we may focus on the part within the parentheses. The calculation in (2.38) suggests that the non-saturating surface potential, the gate stack charges, and the work-function difference may have the same order of weight; in the present example, the three factors degrading the drain current roughly 20, 20, and 10 %, respectively, such that the total degradation is 50 %, i.e. the drain current is reduced to 50 % of its ideal value by these three factors in addition to the reduction by the degraded channel mobility. The calculations illustrated by (2.38) show that even for moderate drain voltages (such as $V_D = 0.4$ V), the drain current versus the drain voltage relation becomes significantly more non-linear for the high-k gate stacks, because the term V_D inside the parenthesis in (2.37) becomes a more significant fraction of the rest of the terms inside the parenthesis.

There are some experimental results [27–31] on the degradation of the channel parameters I_D , g_D , and g_m in high-k MOSFETs by the non-ideal factors of $\Delta\phi_{s,inv}$, $Q_{di,gsc}$, and ϕ_{MS} which support the basic suggestions of (2.37) and (2.38). High pressure annealing of Si/Hf-silicate/HfAlO/TiN gate stacks in pure H_2 was reported to bring down the interface state density by a factor of 2 (from 12 to $6 \times 10^{10} \text{ cm}^{-2}$) and enhance the drain current and the transconductance by 10–15 % [27]. Fluorine treatment and GeO_2 passivation of HfO_2/Ge gate stacks was observed to reduce the interface trap density from 4 to $1 \times 10^{12} \text{ cm}^{-2} \text{ V}^{-1}$, while enhancing the drain current by 18 % [28]. Fluorine incorporation into HfO_2/InP and $HfO_2/In_{0.53}Ga_{0.47}As$ gate stacks was reported to improve the drain current and the transconductance and these improvements were attributed to a reduction in the fixed charge and the interface trap density [29]. Passivation of HfO_2/InP gate stacks by an intermediate Al_2O_3 layer was observed to enhance the drain current 2.5 times and the transconductance 4 times while reducing the interface trap density [30]. Fluorine treatment of $InP/Al_2O_3/HfO_2/TaN$ gate stacks was found to increase the drain current 100 %, the transconductance 50 %, and the channel mobility 56 %; this enhancement was attributed to a reduction in the gate stack charge [31].

2.6.1.4 Channel Conductance and Trans-conductance

The channel conductance may be expressed as, cf. (2.35):

$$g_D = \frac{\partial I_D}{\partial V_D} = \frac{W}{L} \mu_{ch} C_{di} \left[\begin{array}{l} (V_G - \varphi_{s,inv,0} - V_{di,gsc} - \phi_{MS,p}) \\ - 2/3 \frac{\sqrt{2q\epsilon_s N_A}}{C_{di}} 3/2 \sqrt{\varphi_{s,inv,0} + V_D - V_D} \end{array} \right]$$

or,

$$g_D = \frac{W}{L} \mu_{ch} C_{di} [(V_G - \varphi_{s,inv,th,0} - \Delta\varphi_{s,inv,0} - V_{di,gsc} - \phi_{MS,p}) - V_D - V_{di,sc,L}] \quad (2.39)$$

The term $V_{di,sc,L}$ represents the gate stack potential at the drain ($y = L$) due to the ionized dopant charge; it is a function of both the gate and the drain voltages. Equation (2.39) may be compared with its classical formulation, cf. (2.15):

$$g_D = \frac{W}{L} \mu_{ch} C_{di} [(V_G - \varphi_{s,inv,th,0}) - V_D - V_{di,sc,inv,L}] \quad (2.40)$$

The quantity $\varphi_{s,inv,th,0}$ is a constant. The term $V_{di,sc,inv,L}$ represents the gate stack potential due to the ionized dopant charge at the drain ($y = L$) at the onset of strong inversion; it is a function of the drain voltage but not of the gate voltage. The channel conductance may be normalized by its ideal value and expressed as, cf. (2.39), (2.40):

$$\begin{aligned} g_{D,norm} &= \frac{g_{D,high-k}}{g_{D,ideal}} \\ &= \frac{\mu_{ch,high-k}}{\mu_{ch,ideal}} \frac{V_G - \varphi_{s,inv,th,0} - \Delta\varphi_{s,inv,0} - V_{di,gsc} - \phi_{MS,p} - V_D - V_{di,sc,L}}{V_G - \varphi_{s,inv,th,0} - V_D - V_{di,sc,inv,L}} \end{aligned} \quad (2.41)$$

The normalized channel conductance has many parameters in common with the normalized drain current, cf. (2.37) and (2.41); however, g_D is degraded more than I_D , as may be illustrated by making a rough numerical estimate of $g_{D,norm}$ assuming the same values of the parameters, selected for estimating the normalized drain current, cf. (2.38):

$$\begin{aligned} g_{D,norm} &= \frac{\mu_{ch,high-k}}{\mu_{ch,ideal}} \frac{2.2 - 0.9 - 0.2 - 0.2 - 0.1 - 0.4 - 0.12}{2.2 - 0.9 - 0.4 - 0.11} \\ &= \frac{0.28}{0.79} \frac{\mu_{ch,high-k}}{\mu_{ch,ideal}} = 0.35 \frac{\mu_{ch,high-k}}{\mu_{ch,ideal}} \end{aligned} \quad (2.42)$$

As (2.38) and (2.42) indicate, the estimated degradation in channel conductance is nearly 50 % higher than that in the drain current.

The transconductance may be expressed as, cf. (2.35):

$$g_m = \frac{\partial I_D}{\partial V_G} = \frac{W}{L} \mu_{ch} C_{di} \left[\left(1 - \frac{\partial \varphi_{s,inv,0}}{\partial V_G} - \frac{\partial V_{di,gsc}}{\partial V_G} \right) V_D - \frac{2}{3} \frac{\sqrt{2q\epsilon_s N_A}}{C_{di}} \frac{3}{2} (\sqrt{\varphi_{s,inv,0} + V_D} - \sqrt{\varphi_{s,inv,0}}) \frac{\partial \varphi_{s,inv,0}}{\partial V_G} \right] \quad (2.43)$$

or:

$$g_m = \frac{W}{L} \mu_{ch} C_{di} \left[\left(1 - \frac{\partial \varphi_{s,inv,0}}{\partial V_G} - \frac{\partial V_{di,gsc}}{\partial V_G} \right) V_D - (V_{di,sc,L} - V_{di,sc,0}) \frac{\partial \varphi_{s,inv,0}}{\partial V_G} \right] \quad (2.44)$$

$V_{di,sc,0}$ is the gate stack potential at the source ($y = 0$) due to the ionized dopant charge. Equation (2.44) may be compared with its classical counterpart as expressed below, cf. (2.36):

$$g_m = \frac{W}{L} \mu_{ch} C_{di} V_D \quad (2.45)$$

The transconductance normalized by its ideal value may be expressed as, cf. (2.44) and (2.45):

$$\begin{aligned} g_{m,norm} &= \frac{g_{m,high-k}}{g_{m,ideal}} \\ &= \frac{\mu_{ch,high-k}}{\mu_{ch,ideal}} \left(1 - \frac{\partial \varphi_{s,inv,0}}{\partial V_G} - \frac{\partial V_{di,gsc}}{\partial V_G} - \frac{(V_{di,sc,L} - V_{di,sc,0})}{V_D} \frac{\partial \varphi_{s,inv,0}}{\partial V_G} \right) \end{aligned} \quad (2.46)$$

The equation for the transconductance is clearly different in nature from those for the drain current and the channel conductance, cf. (2.46), (2.41), (2.37); namely, (2.46) contains differentials (all of which are fractions) while (2.41) and (2.37) do not. Moreover, the transconductance is not degraded by the work function difference anomaly. It is important to note the strong dependence of the transconductance degradation on the gate voltage. The rate of change of the surface potential with respect to the gate voltage, $\partial \varphi_{s,inv,0} / \partial V_G$, can be obtained from an experimental $\varphi_s(V_G)$ plot, cf. Fig. 2.15. In Fig. 2.15, $\partial \varphi_{s,inv,0} / \partial V_G$ for both the high-k gate stacks is about 0.25 in strong inversion, while it is only 0.08 for the single SiO₂ gate dielectric. This strongly illustrates how important the non-saturating inversion surface potential is in the case of the high-k gate stacks. Similar to the exercise in the case of the drain current and the channel conductance, cf. (2.38) and (2.42), a numerical estimate could be made for the normalized transconductance:

$$\begin{aligned}
g_{m,norm} &= \frac{\mu_{ch,high-k}}{\mu_{ch,ideal}} \left(1 - \frac{\partial \varphi_{s,inv,0}}{\partial V_G} - \frac{\partial V_{di,gsc}}{\partial V_G} - \frac{0.01}{0.40} \frac{\partial \varphi_{s,inv,0}}{\partial V_G} \right) \\
&= \frac{\mu_{ch,high-k}}{\mu_{ch,ideal}} \left(1 - 1.025 \frac{\partial \varphi_{s,inv,0}}{\partial V_G} - \frac{\partial V_{di,gsc}}{\partial V_G} \right) \\
&= \frac{\mu_{ch,high-k}}{\mu_{ch,ideal}} \left(1 - 1.025 \times 0.25 - \frac{\partial V_{di,gsc}}{\partial V_G} \right) \\
&= \frac{\mu_{ch,high-k}}{\mu_{ch,ideal}} \left(1 - 0.26 - \frac{\partial V_{di,gsc}}{\partial V_G} \right) = \frac{\mu_{ch,high-k}}{\mu_{ch,ideal}} \left(0.74 - \frac{\partial V_{di,gsc}}{\partial V_G} \right)
\end{aligned} \tag{2.47}$$

The rate of change in the gate stack potential due to the gate stack charges, $V_{di,gsc}$, with respect to the gate voltage V_G will depend upon how the trap charge inside the gate stack will change and this will be determined by the trap energy levels and the wave function penetration into the gate stack.

Equations (2.38), (2.42), and (2.47) suggest that the degradation is most severe in the case of the channel conductance, followed by the drain current, and then the transconductance.

2.6.1.5 Factors Attenuating Channel Parameters

We may recapitulate our analysis in [Sect. 2.5](#) and [Sect. 2.6.1](#) to list all the factors which force the drain current to attenuate. Ideally, as already stated in [Sect. 2.5.1](#), the entire gate voltage should be spent in enhancing the channel conductivity equally everywhere in the channel irrespective of the y position. The factors which each consumes and wastes a part of the gate voltage are:

1. The drain voltage V_D . The drain voltage causes the voltage across the gate stack to reduce; the amount of reduction depends upon the position along the direction y —at the drain, the amount of reduction is $= V_D$.
2. The ionized dopant charge Q_{dep} . Ideally, the semiconductor surface charge layer should have only electrons or holes. A part of the gate voltage is wasted in sustaining the dopant charge; this effect is more significant for large drain voltages.
3. Work function difference ϕ_{MS} . Ideally, the semiconductor–metal work function difference ϕ_{MS} should be zero. A part of the gate voltage is directly wasted in neutralizing the non-zero work function difference.
4. Gate stack charge $Q_{di,gsc}$. The gate dielectric should be an ideal one; in other words, the gate stack should have only ideal gate dielectrics as constituents, in which the charge density should be zero. A significant part of the gate voltage is wasted in supporting the gate stack charge $Q_{di,gsc}$.
5. Non-saturating inversion surface potential $\Delta\varphi_{s,inv}$. Ideally, the inversion surface potential should remain frozen at its value at the onset of strong inversion

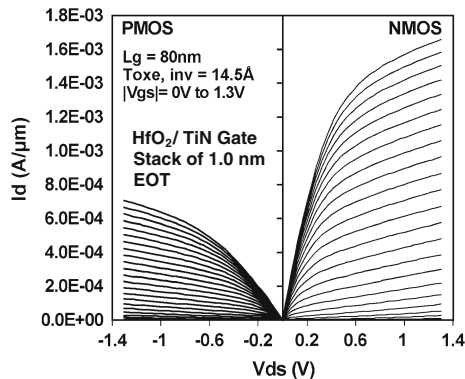


Fig. 2.16 Drain current versus drain voltage characteristics of n-channel and p-channel MOSFETs with HfO_2/TiN gate stack. The EOT was 1.0 nm, whereas the CET was 1.45 nm. The channel length was 80 nm. The gate voltage varied between 0 and 1.3 V. I_D for the n-channel was 1.66 mA/ μm and for the p-channel was 0.71 mA/ μm at a V_D of 1.3 V. Adapted from [32]

$\varphi_{s,\text{inv},\text{th}}$. The excess inversion surface potential $\Delta\varphi_{s,\text{inv}}$ reflects several factors and consumes a significant fraction of the gate voltage.

Figure 2.16 illustrates the drain current versus the drain voltage characteristics measured for both NMOSFETs and PMOSFETs with HfO_2/TiN high-k gate stacks [32]: The channel length was 80 nm, EOT was 1.0 nm, CET was 1.45 nm [32]. It should be of interest to compare the characteristics of Fig. 2.16 representing high-k gate stacks with those of Fig. 2.13 representing SiO_2 single gate dielectric; the two sets of characteristics differ in several respects both in the triode as well as in the saturation regimes. In the triode regime, the high-k characteristics deviate more from a linear form than do the single gate SiO_2 characteristics, whereas in the so-called saturation regime, the former are non-saturating. The deviation from linearity manifested in Fig. 2.16 supports the conclusions of Sect. 2.6.1 that the three non-ideal factors—of $\Delta\varphi_{s,\text{inv}}$, $Q_{\text{di,gsc}}$, and ϕ_{MS} —make the $I_D - V_D$ characteristics more non-linear and degrade the channel parameters of high-k gate stacks significantly, particularly when the supply voltage is reduced, as is the case with the decreasing EOT trend.

2.6.2 Composition of the High-k Gate Stack

The high-k gate stack presents a very formidable challenge to effective (in the sense of being accurate and yet practical in use) modeling and representation. Our aim is to evolve realistic energy band diagrams and circuit representations of the high-k gate stack, and modeling of the various high permittivity layers, and traps and charges, which abound in these systems, and the potentials across the various

dielectric layers. Among the important basic issues, which will be discussed in this section, are:

1. Which constituents/components of the gate stack are to be recognized for representation? Each layer with a different permittivity should in principle be represented by a capacitor in series with those of the adjacent layers.
2. Are there chemically graded layers, intentionally or unintentionally in the gate stack? It has been established [33, 34], that across a Si/SiO₂ interface, there exists a 0.3–0.5 nm thick chemically graded (hence, band-gap and permittivity graded) layer. It is possible that chemically graded layers exist also across the intermediate-layer/high- κ -layer interface and even the high- κ -layer/metal interface (cf. Fig. 2.14). How does one represent a graded permittivity layer in the equivalent circuit, which in principle would be an infinite set of capacitors in series?
3. Is there a metal oxide layer formed between the high- κ bulk and the metal electrode?
4. What are the tangible consequences of the very high density of traps in the gate stack layers? The SiO₂ gate dielectric could be represented by a simple dielectric capacitance (plane-parallel capacitor), because there were no significant traps inside, i.e. it could be represented as an ideal dielectric layer. In great contrast, the charging and discharging in the myriad traps, inside the high- κ gate dielectric stack, in principle, need to be represented by a large number of $R_t C_t$ combinations.
5. This brings us to the issue of the Fermi occupancy of the traps and the profile (x-direction-wise) of the pseudo-Fermi function inside the gate dielectric stack. One could argue that in strong accumulation and in strong inversion, it may be possible for many traps inside the high- κ gate dielectric stack, if not all of them, to follow the applied signal, and exchange electrons/holes, either with the silicon bands or with the metal, if the EOT < 1.0 nm.
6. Is there a Schottky barrier at the high- κ /metal interface?

The basic to any treatment of the gate stack, be it the energy band diagram, the charge analysis, the potential profile, or the circuit representation, is the identification or selection of the constituents of the gate stack. As already discussed, no unique identification is possible, and one has to make a choice or compromise between complexity, accuracy, and practicality. An important point in this context is the strong variation in the approach and practice for an effective passivation of the different semiconductor (Si, Ge, GaAs, InGaAs) surfaces. This may result in significantly different gate stacks and metal electrodes on different semiconductor substrates. Our analysis in this section will be based upon the silicon substrate. We may have a composition of the gate stack, as complicated as illustrated in Fig. 2.17. On the silicon substrate, the intermediate layer is typically SiO₂ or a silicon-oxynitride, whereas the most common high- κ layer, currently, is a Hf-based oxide or oxynitride or silicate or aluminate with or without a dopant (e.g. Y, La). The grading of the interfacial layer between the intermediate layer and the high- κ

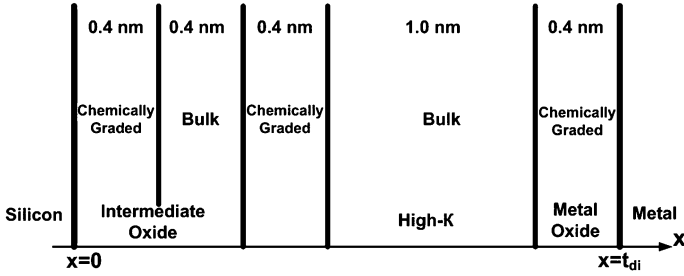


Fig. 2.17 Schematic representation of the five different layers, each with a characteristic permittivity, constituting the high-k gate stack

layer and that between the high-k layer and the gate metal has not been investigated as much as has been the Si-SiO₂ graded interface [33, 34], but graded layers are possible, when one considers atomic, ionic, and inter-layer diffusions, that are likely to occur in the gate stack, due to high concentration gradients [25, 26]. The thickness indicated in Fig. 2.17 for the graded layer at the Si-SiO₂ interface has been established [33, 34], but for the other two graded layers is only indicative.

2.6.3 Energy Profile of the High-k Gate Stack

The representation of Fig. 2.17 would be too complicated to deal with even in terms of the energy band profile, let alone the circuit representation, the representation of the traps, and the electrostatic relations. To arrive at a practical solution, we may consider the following simplifications:

1. The possibility of a metal oxide between the high-k layer and the metal electrode may be ignored. Investigations suggest oxidation of the metal electrode surface and the resultant presence of an oxide to be likely if the metal electrode (e.g. AlN) contains a reactive metal such as Al.
2. Representation of the graded layers may be eliminated to avoid complications too difficult to deal with.
3. The above two simplifications will leave us with two bulk layers—the intermediate layer (e.g. SiO₂ or SiON or a silicate) and a high-k layer (e.g. Hf- or La-silicate or aluminate or oxynitride)—and their three interfaces—Si/IL, IL/high-k, and high-k/metal.
4. The traps and charges at or near the Si/IL interface are comparatively small, as conductance measurements indicate these traps in the device quality MOSFETs to be of the order of 10^{11} cm^{-2} , i.e. orders of magnitude less than the high-k charges and traps. Experimental results will be presented in the later sections.
5. The traps and charges inside the IL are also smaller, as evidence to be presented in the later sections will indicate these to be an order of magnitude smaller than the high-k charges. Experiments on devices with a graded IL indicate these traps and charges to cause voltage drops not exceeding a few tens of mV.

6. The above simplifications will leave us with traps and charges at the IL/high-k interface, the same in the high-k bulk, and in the semiconductor space charge as the dominating ones; these are the traps and charges whose influence on the energy band profile, we will consider.

Figure 2.18 represents a schematic of the simplified high-k gate stack, consisting of two dielectric layers, namely the intermediate layer (IL) and the high-k layer, and three interfaces, namely the silicon/IL interface, the IL/high-k interface, and the high-k/metal interface. Figure 2.18 illustrates the location of the dominant gate stack traps and charges, and identifies the three most important charge locations, which are likely to dominate the electrostatic relations and the carrier transport through the gate stack. It may be borne in mind that the gate stack reliability may not necessarily be dominated by these charge centers. The dominant charges suggested by Fig. 2.18 are:

1. The semiconductor space charge of density Q_{sc} spread over the space charge layer of width W , which has been analyzed in Sect. 2.2.4. This charge will cause significant voltage drop across the gate stack.
2. The traps and charges inside the high-k layer. This charge density is of the order of 10^{13} cm^{-2} , and causes voltage drop of the order of hundreds of mV. The profiles (x-direction variation) of these traps and charges have not been firmly established.
3. The traps and charges at the IL/high-k interface may consist of two entities—an interface dipole and/or interface traps. A dipole, i.e. a positive and a negative layer of equal but opposite charges, of large magnitude at the IL/high-k has been well established by the experimental results [25, 26]. Traps of significant

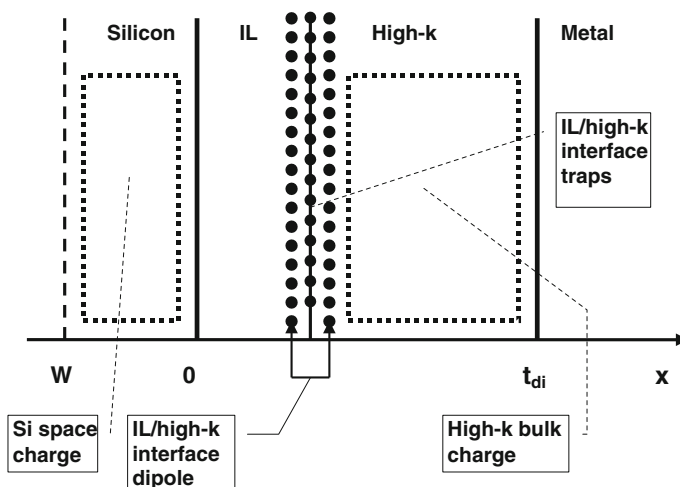


Fig. 2.18 Simplified schematic representation of a high-k gate stack. W is the semiconductor space charge width; t_{di} is the total gate stack physical thickness

density are likely to be present at the IL/high-k interface due to a large mismatch in the chemical bonding.

Several issues need to be considered to render an energy profile representation of even this simplified gate stack.

Image Force Barrier Lowering—When an electron/hole leaves the emitting surface, it experiences an attractive force due to its image charge in the emitter. An attractive force lowers the potential barrier to the electron/hole. Hence, the band offsets and the electron/hole energy barriers are reduced. As the attractive force depends upon the distance between the free carrier and its image charge, the entire gate stack energy profile is modified and the potential energy barrier maximum is shifted from the interface. The image force barrier lowering $\Delta\phi_b$ may be expressed as [2]:

$$\Delta\phi_b = \sqrt{\frac{qE_{\max}}{4\pi\epsilon_{di}}}; \quad x_{\max} = \sqrt{\frac{q}{16\pi\epsilon_{di}E_{\max}}} \quad (2.48)$$

E_{\max} is the maximum electric field, and x_{\max} is the shift in the position of the maximum barrier energy. Equation (2.48) shows that the image force barrier lowering becomes less important in the case of the high-k dielectrics; hence, it may be ignored in the interest of simplification.

Schottky Barrier Formation at the High-k/Metal Interface—As discussed in Sect. 2.2.1, a free carrier exchange across a metal/semiconductor interface leads to the formation of a Schottky barrier; the dipole consists of a charge layer on the metal surface and the space charge layer in the semiconductor sub-surface. A dipole and therefore a Schottky barrier, in principle, cannot form at the metal/ideal-dielectric interface (e.g. at the SiO_2 /metal interface), because an ideal dielectric, or even a near-ideal dielectric, such as the dry thermal SiO_2 , does not have any free carriers to exchange with the metal. A high-k dielectric layer, likewise the dry thermal SiO_2 , has no free carriers in the conduction/valence band, but in strong contrast to an ideal dielectric, has a high density of traps, with which an exchange of free carriers can, in principle, take place with the metal. The effect of such a Schottky barrier on the electrostatic relations and the carrier transport is likely to be muted on account of its proximity to the metal electrode and the high band gap of the high-k layer. Therefore, in the interest of practicality, we will not consider this feature.

Figure 2.14 represents the energy profiles across a p-silicon/ SiO_2 / HfO_2 /TaN MOS capacitor with an intermediate layer of about 1 nm thick SiO_2 , a high-k layer of about 2 nm HfO_2 , and a total EOT of about 1.9 nm. The profile of the vacuum level has been represented. The vacuum level is the hypothetical energy level of an electron in absolute vacuum (Where no other entity exists with which the electron could interact.). It is the highest potential energy an electron can have, and is used as an electron energy reference, i.e. against which other electron energies are compared. All the electrostatic potentials, i.e. the surface potential ϕ_s and the components of the gate stack potential V_{di} , have been marked. It may be noted that the internal vacuum level, being a potential energy, has necessarily to follow the

same profile as the electrostatic potential. The total potential across the gate stack can have different components, depending upon which physical components of the gate stack have been recognized. In the representation of Fig. 2.14, we have recognized three components, namely, the intermediate layer (the dry thermal SiO₂ layer), the high-k layer (the HfO₂ layer), and the dipole at the SiO₂/high-k interface.

An important point that is illustrated in Fig. 2.14 is that the net electric field and the electrostatic potential across any component of the gate stack do not have to be of the same direction or polarity as those obtaining across any other component of the gate stack. In turn, the net electric field and the potential across any gate stack component can result from a number of diverse charges not having the same polarities. The charges, which we have considered in constructing the potential profile of Fig. 2.14, are: the semiconductor space charge Q_{sc} , the bulk high-k charge, the dipole charge, and the charge of traps at the IL/high-k interface; these charges have been illustrated in Fig. 2.18; each of these charges contributes to the net electric field and the net potential across the high-k (HfO₂) layer. As the p-type semiconductor sub-surface is in accumulation, cf. Fig. 2.14, the space charge Q_{sc} consists of holes and is positive; hence the electric field is positive and the surface potential ϕ_s is negative (has value of -0.42 V at a bias of -1.82 V [16], as obtained from the Berglund integral). As we have ignored traps at the Si/SiO₂ interface as well as in the bulk SiO₂ traps, the charge contributing to the field across the SiO₂ layer is only Q_{sc} ; hence, the electric field across the SiO₂ layer is positive and the potential $V_{di,IL}$ is negative (estimated to be about -0.44 V).

The electrostatic picture for the HfO₂ layer or for that matter any other high-k layer, is complicated because of the diverse charges present in its bulk and at its interfaces. The experimental data [16] on the MOS capacitor of Fig. 2.14 (to be presented in more detail later) indicate a total gate stack potential V_{di} of about -1.38 V corresponding to an applied bias of -1.82 V. For the approximations made on the gate stack charges, the total gate stack potential may be expressed as the net sum of the following components:

$$V_{di} = V_{di,IL} + V_{di,dipole} + V_{di,IL/high-k} + V_{di,sc} + V_{di,high-k} \quad (2.49)$$

$V_{di,IL/high-k}$ and $V_{di,high-k}$ are potentials across the HfO₂ layer due to the charge density of the traps at the IL/high-k interface and the high-k bulk charge density, respectively. The potential across SiO₂ layer $V_{di,IL}$ has been estimated to be about -0.44 V. The potential $V_{di,sc}$ across the HfO₂ layer due to the semiconductor space charge Q_{sc} is estimated to be about -0.17 V. This will suggest that the net sum of ($V_{di,dipole} + V_{di,IL/high-k} + V_{di,high-k}$) is about -0.77 V. From the experimental data [16], under the flat-band condition, the potential across the HfO₂ layer due to its bulk charges $V_{di,high-k}$ is estimated to be about $+0.10$ V, and the sum of ($V_{di,dipole} + V_{di,IL/high-k}$) is estimated to be $+0.15$ V. Hence, at flat-band, the sum of ($V_{di,dipole} + V_{di,IL/high-k} + V_{di,high-k}$) is $+0.25$ V, whereas in strong accumulation at an applied bias of -1.82 V, this sum is -0.77 V. This discrepancy can be

reconciled, if all or some of the high- k related charges considered by us vary with the applied bias.

2.6.4 Occupancy of Interface Traps and Bulk Traps in the High- k Gate Stack

The Fermi-Dirac distribution function, in which the Fermi level is the critical parameter, determines the occupancy of an eigenstate. The Fermi level (popular name for the chemical potential, and sometimes mixed up with the Fermi energy, which is defined only at 0 K.) and the law of mass action ($pn = n_i^2$) are thermal equilibrium concepts. The law of mass action is synonymous with a common Fermi level for both electrons and holes. Strictly speaking, thermal equilibrium no longer holds once a bias is applied across the MOS capacitor.

The concept of the quasi-Fermi level (imref—Fermi written in reverse) has no rigorous basis—it is only a practical tool. The electron imref at any location enables us to determine the free electron density at that point, while the hole imref enables us to determine the hole concentration. The deviation of the hole imref from the electron imref represents the scale of the thermal non-equilibrium and the magnitude of the direct current flowing at that point. Notwithstanding its empirical basis, the imref tool has been extensively used inside the semiconductor space charge layer. The occupancy of traps in the dielectric/insulator bulk is a rarely discussed subject; and the extension of the quasi-Fermi level concept to the inner region of a dielectric or insulator is questionable, as there are no free carriers in its conduction/valence band, except in transparent conductors such as $\text{SnO}_2/\text{In}_2\text{O}_3$.

As already mentioned, the high- k dielectric (its bulk and its interfaces) may contain a multitude of traps and charges possibly of different origin and a multitude of electron and hole trap levels; hence it is necessary to know the occupancy of the diverse traps, and the variation of the trap occupancy with the applied bias.

The concept of pseudo-Fermi function and pseudo-Fermi level has been invoked a long time ago to represent the occupancy of a trap inside a dielectric [12], but has rarely been used and developed further. This concept would approximately (i.e. the 0 K approximation) mean that the gate stack traps below the pseudo-Fermi level are occupied by electrons, and are empty if above. In Fig. 2.14, the hole (i.e. the majority carrier) imref in the semiconductor space charge region, and the pseudo-Fermi level in the gate stack have been illustrated. It would be meaningful to apply the concept of the pseudo-Fermi level inside the gate stack, only if the gate stack is thin enough for significant wave function penetration and tunneling. This context leads us to discuss the topic of quantum-mechanical tunneling, which has an important bearing upon a variety of phenomena in and around the high- k gate stack.

2.6.5 Potential Well and Quantum-Mechanical Phenomena

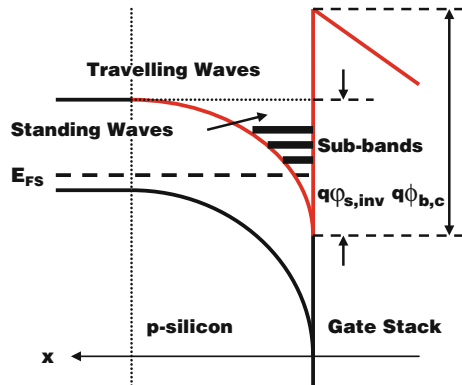
In quantum mechanics, all properties of the electron—its eigen (discrete) energy, the probability of finding it in a given volume of space, its energy bands, its effective mass—are obtained by solving the Schrödinger equation: $H\psi = E\psi$, where the operator H is the Hamiltonian (sum of the potential energy V and the kinetic energy of the electron), E is the eigen-energy, and ψ is the electron wave function. The electron wave function ψ has no direct physical meaning, however, the entity $\psi\psi^*dV$ represents the probability of finding the electron in the infinitesimal volume dV . For no situation in a solid, there is a closed-form solution for the Schrödinger equation, because even in a perfect periodic crystal, the potential energy term is complicated even in its approximated expression. We need to invoke some quantum-mechanical concepts in order to gain some insight into the following phenomena in and around the gate dielectric stack:

Carrier confinement in the potential well of the strong inversion layer (or of the accumulation layer)—Strong inversion or accumulation leads to the formation of a potential well in the semiconductor sub-surface; the profile of the potential well is defined by the (conduction or valence) band bending in the semiconductor, $\phi(x)$, and the band offset at the semiconductor/gate-stack interface, $\phi_{b,c}$ or $\phi_{b,v}$, see Fig. 2.19. An electron or a hole, having a kinetic energy less than the barrier energy $q\phi_s$, is confined in the x -direction to the perimeters of the potential well, and can no longer be represented by a Bloch wave function, i.e. a travelling wave having the periodicity of the lattice, see (2.50):

$$\Psi_{nk}(\mathbf{r}) = u_{nk}(\mathbf{r}) \exp(i\mathbf{k} \cdot \mathbf{r}) \quad (2.50)$$

However, an electron with a kinetic energy exceeding the barrier energy $q\phi_s$ may be represented by a travelling wave, while an electron with a kinetic energy less than $q\phi_s$ will be represented by a standing (or stationary) wave, cf. Fig. 2.19. The latter electrons are localized in the potential well and will be characterized by bound states in the energy sub-bands, cf. Fig. 2.19.

Fig. 2.19 Energy band profile across a p-Si/gate-stack system in strong inversion, illustrating the effects of carrier confinement in the potential well (barrier energy highlighted in red) \gg energy sub-bands, standing waves, etc



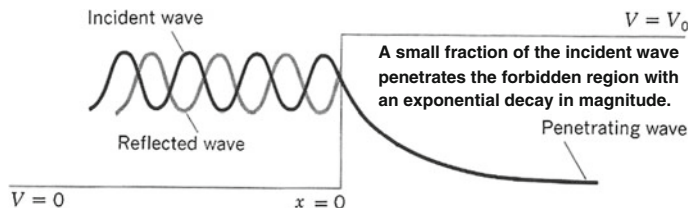


Fig. 2.20 Penetration of a wave, incident at a step potential barrier, into a classically forbidden region. The incident electron energy E is $< V_0$. A small fraction of the incident wave penetrates; the rest is reflected. Adapted from [35]

Carrier wave function penetration into the forbidden potential energy barrier of the gate stack—As the dielectric layers of the gate stack have high band gaps, the gate stack presents a potential energy barrier to the free carriers in the semiconductor and the metal, see Fig. 2.19. In classical physics, an electron of energy E incident at a potential barrier V ($V > E$) will be turned back, i.e. totally reflected, at the so-called classical turning point, cf. Fig. 2.20. In the quantum-mechanical description, unless the potential barrier V is infinite, which is never the case in a gate stack or in any reality, there will exist a finite, however small, probability of finding the electron inside the potential barrier. In other words, the electron wave function penetrates the potential barrier with an exponentially decaying amplitude, and since a wave function exists, the probability $\psi\psi^*dV$ of finding it at a point inside the barrier is finite, cf. Fig. 2.20. To be of some consequence, this point inside the potential barrier, as represented by the gate stack, has to be within a nm or so from the semiconductor or the metal surface. It may be noted that wave function penetration can occur from both the semiconductor as well as from the metal surface.

Tunneling of the free carriers from the semiconductor to the metal through the composite potential barrier of the gate stack and vice versa—Free carrier transport by tunneling across a potential barrier occupies a special place in quantum mechanics, as its quantum mechanical formulation is relatively simple, and, more importantly, it was perhaps the first demonstration of the validity of the quantum mechanical description of matter including the wave-particle duality of an electron. The name of the process under discussion (i.e. tunneling) derives itself from an analogy to a tunnel through a mountain barrier (standing for the forbidden potential barrier). According to the concept of elastic tunneling, a free carrier of energy E incident at a potential barrier of height V ($V > E$) and thickness t , can be transmitted through the potential barrier to an empty or partially empty eigenstate of the same energy on the other side of the barrier, cf. Fig. 2.21. The tunneling transmission coefficient T or the tunneling probability is exponentially dependent upon the product of the barrier thickness and the square root of the excess barrier energy (motive energy) $[V(x) - E]$. Due to its stronger dependence on the barrier thickness, the tunneling probability becomes insignificant for barriers much thicker than a nm. To gain a rough estimate of the transmission coefficient, one could use a rule of thumb: For a motive energy of an eV and a tunneling electron/hole mass of

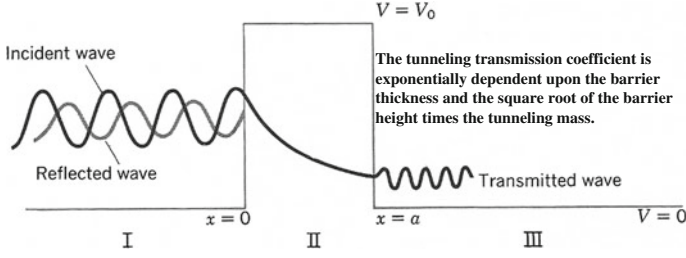


Fig. 2.21 Wave penetration and quantum-mechanical tunneling through a *rectangular barrier* of height V_0 and thickness a , of an electron of energy $E < V_0$, incident at the barrier at $x = 0$ from the *left*. The incident wave is partly transmitted and partly reflected. Adapted from [35]

1.0 m, the transmission coefficient reduces by $1/e$ per each 0.1 nm, such that for 1 nm thick barrier, $T \cong e^{-10}$ if $(V - E)$ is 1 eV, and $T \cong e^{-20}$ if $(V - E)$ is 4 eV.

2.6.5.1 Tunneling Through the Gate Stack

Types of tunneling, relevant for the high- k gate stack, include: direct elastic tunneling (incident and transmitted states are of the same energy), inelastic tunneling (incident and transmitted states are of different energy), trap-assisted tunneling (could be a chain process mediated by several traps), and Fowler-Nordheim tunneling (tunneling into a conduction/valence band in the gate stack). Theoretical treatments exist for direct tunneling through a rectangular potential barrier and other simple barriers in the text books on quantum mechanics [35, 36]. Simplest is the case for a rectangular potential barrier (of height V_0), see Fig. 2.21, for which solution of the one-dimensional, time-independent Schrödinger equation:

$$-\frac{\hbar^2}{2m^*} \Psi'' + (V_0 - E)\Psi = 0, \quad (2.51)$$

yields the following closed-form wave functions for an electron of energy E incident on the barrier from the left:

$$\Psi_I = Ae^{ikx} + Be^{-ikx}, \quad x < 0 \quad (2.52)$$

$$\Psi_{II} = Ce^{\kappa x} + De^{-\kappa x}, \quad 0 < x < a \quad (2.53)$$

$$\Psi_{III} = Fe^{ikx}, \quad x > a \quad (2.54)$$

$$k = \sqrt{\frac{2m^*E}{\hbar^2}}, \quad \kappa = \sqrt{\frac{2m^*}{\hbar^2}(V_0 - E)} \quad (2.55)$$

\hbar is Planck's constant/ 2π , m^* is the tunneling effective mass, A , B , C , D and F are constants, k is the electron wave vector, κ is the wave attenuation constant, and a is the barrier thickness. It is not clear what the tunneling effective mass should be.

First of all, the tunneling process does not involve any electron motion in the usual sense; so Newton's second law of motion and the usual effective mass concept may not apply. Secondly, it is also not clear whether any motion at all is involved in the tunneling process.

Application of the boundary conditions (continuity of the wave function and its derivative at the boundaries $x = 0$ and at $x = a$) yields the following relations among the constants:

$$\begin{aligned} C &= \frac{1}{2\kappa} \{(\kappa + ik)A + (\kappa - ik)B\} = \frac{\kappa + ik}{2\kappa} e^{-(\kappa - ik)a} F \\ D &= \frac{1}{2\kappa} \{(\kappa - ik)A + (\kappa + ik)B\} = \frac{\kappa - ik}{2\kappa} e^{-(\kappa + ik)a} F \end{aligned} \quad (2.56)$$

The profile of the potential barrier presented by the gate stack, $V(x)$, is very different from a rectangular shape. Nevertheless, the relations in (2.52–2.56) could still be applied, if we replace V_0 in (2.55) by the potential energy profile $V(x)$ across the gate stack. For EOT of 1 nm or less, the gate leakage current is very significant and is one of the most important issues for the MOSFET. The dominant mechanism of carrier transport through the gate stack is likely to be some form of tunneling. Hence, the ability to reliably estimate the tunneling current would be very useful. Unfortunately, this ability is seriously compromised by the following factors, among others:

1. The profile of the potential barrier across the gate stack, $V(x)$, cannot be known, because the composition of the layers of the gate stack is complicated by interlayer diffusion and chemical reaction, which in turn determine the energy bands and the tunneling mass.
2. It is not clear what mass one is to use for the free carriers in the tunneling equations. Even if the usual effective mass is applied, this entity is not accurately known for most of the high- k dielectrics.

2.6.5.2 Carrier Confinement in the Strong Inversion and Accumulation Layers

The strong inversion layer, i.e. the MOSFET channel, could be a few nm thick; hence for an electron (hole) inside this potential well, i.e. for $E < V_0 = q\phi_s$, the movement in the x -direction (i.e. perpendicular to the interface) is restricted, although parallel to the surface, the electron movement is free and the 2D (2-dimensional) periodicity of the semiconductor crystal is retained. This restriction in the x direction leads to a quantization of the eigenstates of the conduction or the valence band, ultimately resulting in sub-bands inside the potential well, separated by regions of forbidden energy; cf. Fig. 2.19; however, for electron energies higher than $V_0 = q\phi_s$, the usual distribution of states in the band remains basically

unaltered. In other words, an electron ($E > V_0 = q\phi_s$) inside the potential well is less delocalized than the channel electrons with $E > V_0 = q\phi_s$, which are still treated as a free electron gas. The Schrödinger equation for an electron with $E < V_0 = q\phi_s$ could be expressed as:

$$\left[-\frac{\hbar^2}{2} \left(\frac{1}{m_x} \frac{\partial^2}{\partial x^2} + \frac{1}{m_y} \frac{\partial^2}{\partial y^2} + \frac{1}{m_z} \frac{\partial^2}{\partial z^2} \right) + V(x) \right] \Psi(\mathbf{r}) = E\Psi(\mathbf{r}) \quad (2.57)$$

where $V(x)$ is the electrostatic potential energy in the strong inversion or the accumulation layer and is a function of x only. A solution of (2.57) could be expressed as:

$$\Psi(\mathbf{r}) = \phi_i(x) \exp(ik_x y + ik_z z) \quad (2.58)$$

The x component of the bound electron is obtained by solving the equation:

$$\left(-\frac{\hbar^2}{2m_x} \frac{\partial^2}{\partial x^2} - V(x) \right) \phi_i(x) = \varepsilon \phi_i(x) \quad (2.59)$$

For the electrostatic potential energy $V(x)$, the boundary conditions are: $V(x = 0) = -q\phi_s$ and $V(x = \infty) = 0$. The electrostatic potential energy $V(x) = -q\phi(x)$, and as demonstrated in Sect. 2.2.4, the electrostatic potential $\phi(x)$ (i.e. the band bending) are obtained by solving the Poisson equation, in which the net charge density $\rho(x)$ will take the expression:

$$\rho(x) = q \left(-\sum_i n_i |\phi_i(x)|^2 + \sum N_D^+ - \sum N_A^- \right) \quad (2.60)$$

where n_i is the electron density having the i -th eigenenergy ε_i . The two relations (2.59) and (2.60) are coupled requiring a self-consistent solution of the Schrödinger and the Poisson equations [37, 38]. The treatment of the quantization of the accumulation layer is similar to that of the strong inversion layer, except the following item. A weak inversion layer and a depletion layer separate the neutral region from the strong inversion layer, which is not the case in accumulation; consequently, the free electrons (or holes) have also to be included in the expression for the space charge density $\rho(x)$ in addition to those in the sub-bands.

Following are among the significant consequences of the quantization of the strong inversion and the accumulation layers:

1. Quantization significantly alters the electron (hole) density profile $n(x)/p(x)$ at the semiconductor surface in strong inversion and accumulation, as illustrated in Fig. 2.22. In the case of the 3-dimensional (3D) Bloch wave representation, i.e. in the classical analysis, the electron (hole) density is determined solely by the energy separation between the band-edge (E_c or E_v) and the semiconductor Fermi level E_{FS} at the interface ($x = 0$); hence, the electron (hole) density peaks at the interface, cf. Fig. 2.22. However, in the 2D representation, the

wave function is required to vanish at its nodes, see Fig. 2.23, which includes the interface and the other perimeter of the potential well. This approach is similar to the text-book treatment of a particle in a box in one dimension [39], where infinite potential barriers are assumed at the boundaries of the box, cf. Fig. 2.24. The same assumptions of infinite potential barriers have been made in most of the treatments of carrier confinement in strong inversion and accumulation layers [37, 40]. In the 2D representation, the electron density does not peak at the interface, but away from the interface and inside the potential well, cf. Fig. 2.22.

2. The free carrier charge density of the MOSFET channel (strong inversion layer) or the accumulation layer is reduced, resulting in a lower space charge density Q_{sc} and therefore a lower space charge capacitance density C_{sc} for the same value of the surface potential (or band-bending). The gate dielectric capacitance density C_{di} is also said to effectively reduce on the basis of the following argument: The gate dielectric separates equal and opposite charges on its two sides. Electric field lines emanate from the metal surface and terminate in the semiconductor space charge layer or the vice versa. The effective separation between the metal surface charges and the semiconductor space charge layer charges increase, see Fig. 2.22, leading to an effective increase in the capacitive dielectric thickness. As a result, the total saturated capacitance of the MOS structure, C , is reduced.
3. The effective semiconductor band-gap is increased in strong inversion and in accumulation, as the first sub-band ε_0 lies above/below the band edge E_c/E_v by a significant amount, cf. $\Delta\varepsilon$ in Fig. 2.23.

Fig. 2.22 Electron density profile $n(x)$ in the strong inversion layer for a silicon substrate at 150 K with (100) orientation, acceptor density of $1.5 \times 10^{16} \text{ cm}^{-3}$ and a total electron density of 10^{12} cm^{-2} . The *broken line* represents the contribution to $n(x)$ of the lowest sub-band alone. Adapted from [37]

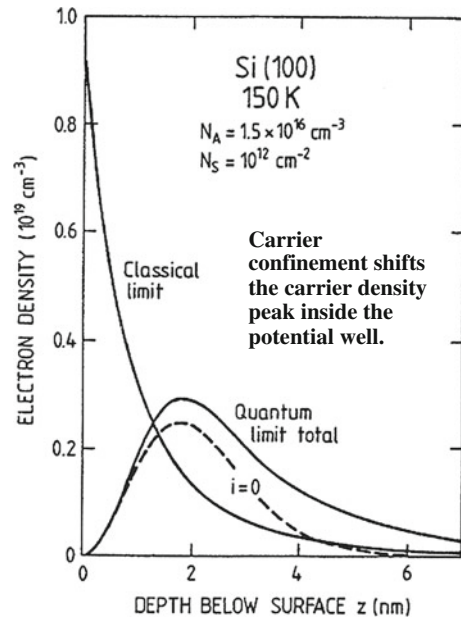


Fig. 2.23 Schematic illustration of the electron wave functions ϕ_0 and ϕ_1 for the lowest two sub-bands ε_0 and ε_1 , for a triangular potential well approximation of the strong inversion or the accumulation layer. Adapted from [38]

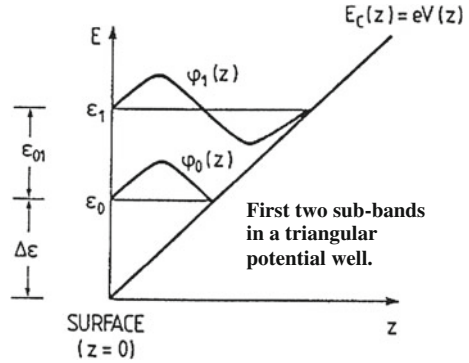
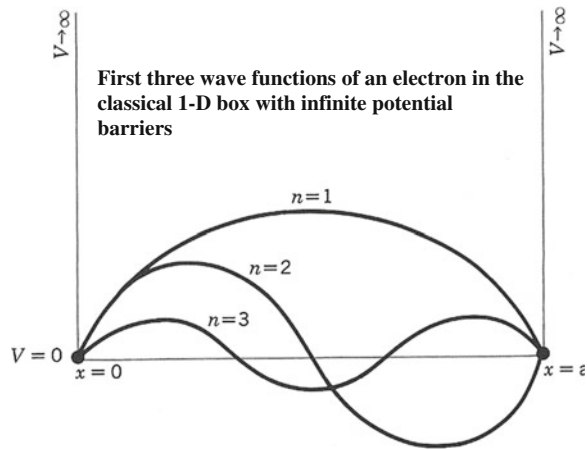


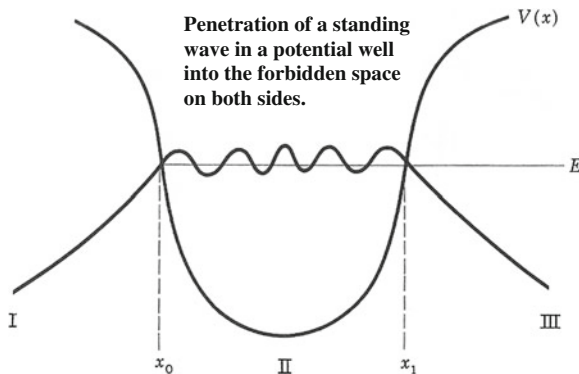
Fig. 2.24 Schematic representation of the first three wave functions for an electron confined to a one-dimensional box. Adapted from [35]



There exists experimental evidence for the validity of the 2D representation of the strong inversion and the accumulation layers and the resultant sub-bands, from infrared absorption and magneto-conductance experiments. However, there are good reasons to believe that many of the quantization and 2D treatments [37, 41], which appear to be very popular, overestimate the effects of the carrier confinement. One reason for the popularity could be that these estimates or calculations predict an EOT, which is significantly lower than what the physical dimensions would allow; *this—much lower estimate of the EOT—would appear to be a success and a progress for realizing the target set by the ITRS roadmap*. The carrier confinement effects may be overestimated for the following reasons:

1. *Electron wave function penetration*—As mentioned already, the potential barrier was assumed to be infinite at the boundaries of the potential well, see Fig. 2.24, to simplify the mathematical treatment. For the high- k gate stack with much lower band offsets than SiO_2 , and, generally, with much smaller effective mass, this assumption is hard to justify [It may be noted that the wave

Fig. 2.25 Penetration of the bound state E (standing wave) in a potential well into the classically forbidden regions outside the potential well (exponentially attenuating wave). Adapted from [35]



attenuation constant depends upon the product of the effective mass and the barrier height, cf. (2.55)]. Penetration of the bound states (in classically allowed regions) into both sides of the classically forbidden regions is a text book problem in quantum mechanics, see Fig. 2.25. However, it is only recently, that penetration of the standing waves in the strong inversion or the accumulation layer into the gate stack has been looked into [41]. Calculations indicate significant effect of the wave function penetration in enhancing the electrical oxide thickness: by as much as 0.33 nm or more [42]. Major problems in correctly estimating the extent of wave function penetration into the gate stack include the unknown potential barrier profile in the gate stack and the corresponding tunneling mass, as mentioned in Sect. 2.6.5.1.

2. *Metal induced states*—Another factor which may dilute the carrier confinement effect is the possibility of metal induced states in the semiconductor sub-surface, when EOT is < 1 nm. A metal-induced gap state is an old concept [43], which was invoked in the case of the metal–semiconductor interface. In fact, when EOT is very small, interference between wave functions of carriers at the semiconductor surface and at the metal surface may be a significant possibility; such an interaction may significantly alter the properties of the gate stack.
3. *Very large tunneling currents*—For EOT, say = 0.5 nm, the tunneling current may be as large as a significant fraction of the drain current, and the tunneling time (If that concept is valid) may be of the order of the channel traversing (transit) time and also the semiconductor relaxation time. In such a case, it is not known what the Fermi occupancy of the sub-bands would be and which Fermi level—the semiconductor or the metal—would apply to these sub-bands. The large tunneling current itself will dilute the carrier confinement phenomenon.

2.6.5.3 Wave Function Penetration into the Gate Stack

Penetration of an incident electron wave with energy E into a region of higher potential energy V_0 is classically forbidden, but is quantum-mechanically allowed, as illustrated in Fig. 2.20. It does not matter quantum-mechanically, if the barrier is infinite in thickness, as Fig. 2.20 would suggest. However, there would be no penetration if the barrier energy is infinite. As already outlined, the concept of wave-function penetration is almost as old as quantum mechanics itself. If the wave function ψ exists inside the potential barrier, $\psi\psi^*dV$ would be finite, howsoever small it might be; therefore there would be a finite probability of finding the electron inside the barrier. However, it does not appear that electron/hole traps were contemplated at that time to exist inside the forbidden potential barrier; for, such an entity would also require the corresponding trap energy level (i.e. an eigenstate) to exist. If there are no allowed energies inside the potential barrier, it is not clear in the classical treatment of wave function penetration, how one would find the electron inside the barrier. The issue of traps, the occupancy of these traps, and the related pseudo-Fermi function was investigated many years later [12], but this lone treatment of these issues was not followed subsequently. For a potential barrier profile, where the potential energy is a function of x , it would follow from (2.55), that the electron density attenuation $A = |\psi|^2$ at a distance x from the point of incidence would be given by:

$$A = \exp -2 \sqrt{\frac{2m^*}{\hbar^2} \int_0^x \sqrt{\phi(x)} dx} \quad (2.61)$$

where ϕ is the barrier height.

2.6.6 Trap Time Constant

As already outlined, the high- k gate stack is beset with a high density of traps inside its bulk layers and at the interfaces between the gate stack layers, cf. Fig. 2.26. Figure 2.26 represents the energy band diagram of a high- k gate stack on a p-Si at the onset of strong inversion. This diagram represents schematically the traps at various locations inside the gate stack—at the Si-IL interface, inside the IL layer, at the IL/high- k interface, inside the high- k layer, and at the high- k /metal interface. Indicated in this diagram are the physical thicknesses of the various layers of the gate stack and the different components of the gate stack potentials. The charges in the gate stack, particularly when these are high as inside the high- k layer, will make the potential profile non-linear; for the sake of simplicity, this point has been ignored in Fig. 2.26, and the indicated potential profiles are linear. An important feature illustrated in Fig. 2.26 is the pseudo-Fermi level, cf. Fig. 2.14; the pseudo-Fermi level is meant to indicate the trap occupancy function. As indicated in Fig. 2.26, the pseudo-Fermi function and the trap

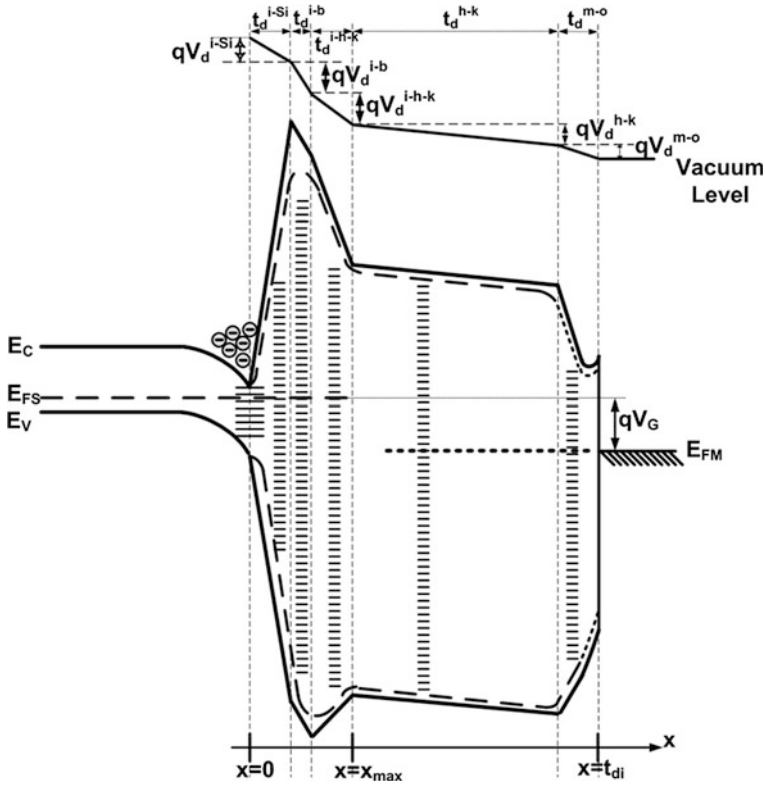


Fig. 2.26 Energy band diagram of an MOS structure with a high- κ gate stack, consisting of bulk intermediate oxide and bulk high- κ layers and chemically graded layers at the three interfaces, illustrating the effects of the graded band-gaps and the electric field. i-Si and i-h-k are respectively the transition layers between Si and IL and IL and high- κ layers. i-b and h-k are respectively the bulk IL and high- κ layers. m-o is the transition metal oxide layer between the high- κ layer and metal electrode. The effect of the charges in the gate dielectric layers, on the potential profile (i.e. variation of the potential along direction x), has not been represented (The actual potential profile will be non-linear). The broken profile schematically represents the effect of the image force on the potential and the energy barrier profile. Figure 2.26 illustrates the formation of a Schottky barrier at the high- κ /metal interface

occupancy are given by the semiconductor majority carrier imref $E_{FS,h}$ up to a distance x_{max} into the gate stack, whereas in the rest of the gate stack it is given by the metal Fermi level E_{FM} .

The gate stack traps are likely to capture and emit electrons or holes, depending upon the bias conditions and the signal frequency. How deep into the gate stack and how many of these traps exchange free carriers with either the semiconductor substrate or the metal electrode, will depend upon the potential barrier profile of the gate stack, including the barrier energy and the physical thickness. For a low EOT, such as 0.5 nm, it is in principle possible for traps at all locations inside the gate stack to communicate either with the semiconductor energy bands or with the

metal energy bands, cf. Fig. 2.26. There may be traps of different physical and chemical nature with their characteristic capture and emission probabilities and relaxation times. The capture or emission time constant of a trap may carry its signature or reflect its identity.

The electron or the hole capture (electron/hole emission is a different process from electron/hole capture) is a complex process, and there is no comprehensive theory for the capture probability of a trap inside a forbidden region (dielectric or an insulator). In a simple formulation, the capture probability is equated to $v_e\sigma_e$ or $v_h\sigma_h$ (v_e/v_h is thermal velocity, and σ_e/σ_h is capture cross-section for electron/hole), which has the unit of cm^3/s . The capture cross-section remains an ambiguous concept; it hides our inability to formulate a clear set of relations for the capture process. Experimental results [44] suggest that the capture cross section can vary over many orders of magnitude, such as from 10^{-12} to 10^{-18} cm^2 . Can one explain such an enormous variation of the capture cross-section? The usual explanation offered is that scattering by Coulomb attraction entails the largest and by Coulomb repulsion the smallest capture cross-sections.

The capture cross-section of a trap would likely depend upon its neighborhood, which may change strongly along the x direction inside the gate stack. So, the capture cross-section may change strongly with the trap location inside the gate stack, x_t ; it may also be strongly dependent upon the trap energy (Modeling [45] indicates, even for one kind of defect, for example for oxygen vacancies, trap levels at different energies with different effective charges.).

In the simplest formulation, one can argue that the capture time, for a trap at the location x_t inside the gate stack, will be inversely proportional both to the density of the free carriers (electron or hole) at x_t and also to the capture probability, cf. Fig. 2.27. The time for a hole capture by the trap located at x_t is then given by:

$$\tau_t^h(x_t) = \frac{1}{v_h\sigma_h p(x_t)} \quad (2.62)$$

where $p(x_t)$ is the hole density at the trap's location.

Since the wave-function of an electron or a hole at the silicon surface (i.e. $x = 0$) can penetrate the potential barrier presented to it by the gate stack, there is a non-zero probability of finding the electron/hole at any trap location x_t , cf. Fig. 2.27. The hole density at x_t is determined by the electron wave function attenuation constant κ_t [12]:

$$p(x_t) = p_s e^{-2\kappa_h x_t}; \quad \kappa_h x_t = \sqrt{\frac{2m_h}{\hbar} \int_0^{x_t} \sqrt{\phi_h(x)} dx} \quad (2.63)$$

where p_s is the hole density at the Si surface, m_h^* is the effective tunneling mass of holes, and $\phi_h(x)$ is the potential energy barrier profile for holes in the gate stack, as defined by the valence band edge, the electric field, and perhaps the image force barrier lowering.

depth, x_{\max} , such that all traps in the gate stack in the range of 0 and x_{\max} would follow the applied small signal. This maximum penetration depth x_{\max} would be higher for a lower signal frequency. Further, as the semiconductor surface carrier density increases with increasing accumulation or strong inversion, x_{\max} would be strongly bias dependant.

To get a feel for this maximum penetration depth x_{\max} , let us consider a 100 kHz signal and a surface hole density of 10^{20} cm^{-3} (This magnitude of hole density would represent a p-type semiconductor in deep accumulation or a n-type semiconductor in deep inversion.). For the traps to follow the applied signal, the condition $\omega (=2\pi f) \ll (\tau_t^h)^{-1}$ has to be fulfilled. Assumption of a hole capture cross-section of 10^{-15} cm^2 , and a hole density of 10^{16} cm^{-3} at the trap location x_t , and a hole thermal velocity of about 10^7 cm/s at 300 K, leads to a trap time-constant of 10 ns, cf. (2.62). Hence, the traps should be able to follow the 100 kHz signal under these conditions. So, for a 100 kHz signal, and a rectangular barrier of 2 eV and $m_h^* = 0.18 m$, according to (2.63)

$$x_{\max} = \frac{\ln \frac{p_s}{p}}{2\kappa_h} = 0.16 \ln \frac{10^{20}}{10^{16}} \text{ nm} = 1.47 \text{ nm}$$

Similarly, for a 10 kHz signal, x_{\max} would be 1.84 nm, assuming $(2\kappa_h)^{-1} = 0.16 \text{ nm}$. An x_{\max} of 1.47 or 1.84 nm would mean that, in the case of high-k gate stacks with EOT = 1.0 nm or less, all traps in the gate stack may follow the 100/10 kHz signal, either through communication with the semiconductor energy bands or through the same with the metal electrode energy bands, cf. Figs. 2.14 and 2.19.

In the case of the ultrathin gate stacks, it is not necessary that the traps inside the gate stack exchange free carriers only with the semiconductor surface. Free carriers are available on the metal surface as well; hence free carriers on the metal surface will compete with the free carriers on the semiconductor surface to fill/empty (i.e. charge/discharge) the gate stack traps. Which exchange will dominate will depend upon the carrier density at the two surfaces and the rate of tunneling. What this means is that the quasi-Fermi occupancy of the traps between $x = 0$ and $x = x_{\max}$ will be given by the semiconductor quasi-Fermi level E_{FS} , while the quasi-Fermi occupancy of traps in the range of $x = x_{\max}$ and $x = t_{di}$ would be given by the metal E_{FM} , as illustrated in Fig. 2.26. In other words, depending on the bias, a part of the gate stack is in quasi-equilibrium with the semiconductor surface, and the rest of the gate stack is in quasi-equilibrium with the metal surface. As the semiconductor surface carrier density is strongly bias-dependent, while the metal surface carrier density is constant (depends on the metal), x_{\max} will be bias-dependent. In deep accumulation or in deep inversion, x_{\max} is likely to be in the vicinity of the plane interfacing with the intermediate layer and the high-k layer, as has been indicated in Fig. 2.26. Consequently, the bulk semiconductor imref E_{FS} and the pseudo-Fermi function will extend into the high-k gate stack up to a distance x_{\max} from the silicon surface, while in the rest of the gate stack, the Fermi occupancy will be given by the metal Fermi level E_{FM} (cf. Figs. 2.26, 2.27).

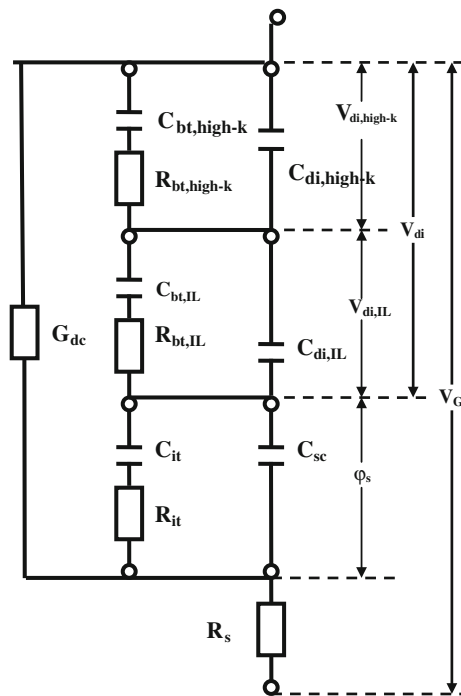
2.7 Nature of Traps and Charges in the High-k Gate Stack

Gate stack traps may be defined as electrically active defects in the gate stack which manifest themselves by capturing or emitting electrons or holes. The traps may be characterized by the localized trap level (an allowed state), its capture and emission cross-section, and its charged state. It is not necessary for the trap energy level to be located inside the band-gap; the trap levels could be located inside the allowed energy bands of the host dielectric. It is not uncommon to come across a perception that the trap states are required to be inside the band-gap. The only real difference is that if the trap state is inside an allowed band, then it supplements the allowed states of the host, whereas if it is inside the band-gap, then it exists where the host has a zero density of states. The charge of a trap may or may not vary with the applied gate voltage; if it does not, then it will act as a fixed charge, and will not contribute to any trap capacitance; if it charges or discharges with the gate voltage, then a trap capacitance will result. As analyzed in [Sect. 2.6.4](#), whether charging or discharging happens at the trap will depend upon whether the trap occupancy changes with the gate voltage.

An unusually high trap density inside the gate stack, as currently obtains even in the device quality high-k gate stacks, has serious implications which have not received much attention so far. We have already made a clear distinction between a dielectric or a dielectric stack, which is free of charges inside, and the high-k gate stack in reality, whose capacitance is not a dielectric capacitance, but is a space charge capacitance. We may recapitulate that a near-ideal dielectric like the SiO_2 has no significant charges inside it with the result that the electric field is constant and the potential is a linear function of distance inside it. As discussed in [Sect. 2.6.6](#), it is possible for all the traps inside the gate stack to communicate with the semiconductor or the metal surface and exchange electrons or holes (i.e. charge or discharge), if the EOT is small (say <1.0 nm), the frequency is 100 kHz or even 1 MHz, and the device is biased in strong accumulation or inversion, cf. [Figs. 2.26](#) and [2.27](#). In such a situation, as we will see in detail in later sections, the trap capacitance (due to trap charging or discharging) will add to the dielectric capacitance of the gate stack, yielding a value of C_{di} higher than what would obtain in the GHz range—the operating frequency of the MOSFET, cf. [Fig. 2.28](#). At GHz frequencies, the traps inside the high-k gate stack are unlikely to exchange electrons or holes with either the semiconductor or the metal surface. In other words, the parameter extraction technique (normally carried out at 1 MHz or below) would yield a higher value of the gate stack capacitance C_{di} than what would obtain during the MOSFET operation and would overestimate its performance. It would also give an erroneously lower value of EOT or CET and induce a false sense of progress on downscaling of the gate stack thickness.

As already mentioned, the topic of surface states and interface states have engaged many researchers over several decades, beginning with William Shockley, Igor Tamm, John Bardeen, Walter Schottky, and Volker Heine. Igor Tamm was perhaps the first to realize the possibility of localized states existing at the

Fig. 2.28 Equivalent circuit diagram for the MOS structure in accumulation and strong inversion, at an intermediate frequency, illustrating the effect of charging/discharging in traps located inside and at the interfaces of the two layers constituting the high- κ gate stack



surface. The classical concepts of Tamm states [46] and Shockley states [47] are generally not invoked in the case of the high- κ gate stacks. Theoretical and even experimental information is incomplete on the nature of traps inside the high- κ gate stack. From common sense one may expect to find different types of intrinsic traps in the different regions of the high- κ gate stack, because the electronic properties of the trap would primarily be decided by its nearest neighborhood:

1. *Si-IL (SiO₂-like) interface*—Transition region from a covalent semiconductor to a primarily covalent insulator, cf. Fig. 2.26. The interface traps, investigated most, experimentally, are those existing at the Si-SiO₂ interface. The experimental information on the characteristics of these traps may be considered both reliable and complete. However, the theoretical understanding and identification of the origin of these traps are incomplete notwithstanding several hypotheses which exist on their origin. From the experimental results, the device grade Si-SiO₂ interface appears to exhibit the following interface state distribution across the Si band-gap: two peaked profiles overlying a U-shaped background. There are good reasons to believe that some of the intrinsic traps at the Si-IL interface may be the P_b center with the structure of Si≡Si₃—amphoteric trap which is a threefold-coordinated Si with a dangling bond [48, 49]. An amphoteric trap can be positively charged, when it has no electron, and is a donor state, neutral, when it has one electron, and negatively charged, when it has two electrons, and is an acceptor state. Experiments suggest that the peak

in the lower Si band-gap (around 0.30–0.35 eV above E_v) represents the donor P_b center, while the peak in the upper Si band-gap (around 0.80–0.85 eV above E_v) represents the acceptor P_b center. Less certain is the origin of the U-shaped background $D_{it}(E)$ distribution. Possible origin could be the tail states of the conduction band and the valence band, like in amorphous silicon [50], and Si-IL strain-induced states. In addition to the intrinsic traps, chemical impurities can generate characteristic traps. Experiments by Kar and Dahlke [51] have convincingly demonstrated that metal impurities from the gate electrode can generate metal-specific trap levels with characteristic capture cross-section. Results presented in many chapters suggest the possibility of high- k cation being present at the Si-IL interface.

2. *IL bulk*—When the semiconductor is Si, the intermediate layer (IL) is SiO_2 or $SiON$. The dry thermal SiO_2 layer normally should be defect-free. But, according to Bersuker, gate stack degradation experiments reveal the precursor defects located in the IL layer, cf. Fig. 2.26, to be the main concern for high- k gate stack reliability, cf. Chap. 8. These may be oxygen vacancies, with trap levels located 2.2–3.5 eV below the silica conduction band [52], induced by the interaction of the intermediate layer with the high- k /metal layers. Stress-induced traps are found to be generated in these precursor defects.
3. *IL/high- k interface*—Transition region from a primarily covalent insulator to a highly ionic insulator with strong mismatch in terms of ionicity (electro-negativity), coordination number, and lattice constant, cf. Fig. 2.26. This multifaceted mismatch can in principle generate a high density of interface traps of various kinds. Unfortunately, it is difficult to access this interface by the interface trap extraction techniques such as the MOS conductance technique and also the charge-pumping technique. However, the limited experimental results suggest a high density of traps at the IL/high- k interface and also indicate that the trap density increases with distance from the Si surface. Experimental results presented in Chaps. 6 and 5 confirm the presence of a strong interface dipole, cf. Fig. 2.14. The origin of this dipole could be the areal oxygen density difference [53] or the electro-negativity difference [54] across the IL/high- k interface.
4. *High- k bulk*—Ionic oxides are always rich in oxygen vacancies; intrinsic defects are likely to be dominated by oxygen vacancies. Extrinsic defects due to diffusion from the semiconductor (e.g. Si) or from the metal are possible. As has been discussed in Chap. 8, in monoclinic hafnia, oxygen vacancies have been shown by ab initio calculations to exist in five charge states, from -2 to $+2$ and may function as electron/hole traps, as well as fixed charges. The IL (SiO_2) layer may be oxygen deficient at the Si/IL interface and also at the IL/Hafnia interface due to the diffusion of oxygen vacancies from the hafnia layer. The results presented in Chap. 8 indicate the Vo^{++} traps to have an energy level about 0.5 eV below the hafnia conduction band, a capture cross-section of 10^{-13} cm^2 , a time constant of 0.5 μs , and an extremely high areal density of 10^{14} cm^{-2} .

5. *High-k/metal interface*—As already mentioned, one does not talk of interface traps at any SiO_2 /metal interface. But a high-k material is not a near-perfect dielectric as SiO_2 is. The concepts of Fermi level pinning and charge neutrality layer have often been invoked recently to analyze the anomalous behavior at the high-k/metal interface. Both of these concepts will not apply unless the interface trap density is extremely high. The moot question remains what could generate these traps. The most unfortunate part is that normally a high trap density should be easier to measure directly. But there is no trap extraction technique available which can be applied on a metal surface. So, all the evidence on these traps is indirect. One possible origin could be tails of the metal wave functions which decay into the high-k layer. In the case of the metal-semiconductor contacts, as these mid-gap states penetrate some distance into the semiconductor, one treats these states as mixtures of the valence and conduction band wave functions with the metal wave functions [55].

2.8 Potentials and Circuit Representations of the High-k Gate Stack

The energy band diagram of Fig. 2.26 illustrates five components of the gate stack potential V_{di} . Three of these potential components are across the three interface regions—Si/IL, IL/high-k, and high-k/metal—in which the chemical composition has been considered to be graded. This may be a more realistic representation, but, the mathematical representation of the electrostatic potential across a layer with a graded composition is too complicated; in particular, the permittivity of such a graded layer will be graded also, which will be difficult to handle. To simplify the equations, we will merge the three interfaces into the two bulk layers—IL and high-k—and consider constant permittivity in each of these layers. Such a less complex situation is illustrated in the energy band profile in Fig. 2.14. With this simplification, the total potential across the gate stack, V_{di} , will have two components, across the IL, $V_{di,IL}$, and across the high-k layer, $V_{di,high-k}$:

$$V_{di} = V_{di,IL} + V_{di,high-k} \quad (2.64)$$

Each of these potentials will depend upon the silicon space charge density Q_{sc} , the charge density in the interface states at the Si-SiO_x interface Q_{it} , the fixed charge density in the proximity of the silicon surface Q_F , the total charges in each of the gate stack layers preceding the layer in question, the charges in the layer itself, the dielectric capacitance of the layer and its permittivity, and the physical thickness of the layer. The potential across the bulk intermediate-oxide layer, $V_{di,IL}$, may be expressed as:

$$V_{di,IL} = -\frac{Q_{sc} + Q_{it} + Q_F}{C_{di,IL}} - \int_0^{t_{di,IL}} \frac{\rho_{IL}(x)dx}{\epsilon_{di,IL}} (t_{di,IL} - x) \quad (2.65)$$

where $C_{di,IL}$ is the dielectric capacitance, ρ_{IL} is the volume charge density, $\epsilon_{di,IL}$ is the permittivity, and $t_{di,IL}$ is the thickness of this layer. The potential across the bulk high-k layer, $V_{di,high-k}$, may be expressed as:

$$V_{di,high-k} = -\frac{Q_{sc} + Q_{it} + Q_F + \int_0^{t_{di,IL}} \rho_{IL}(x)dx}{C_{di,high-k}} - \int_{t_{di,IL}}^{t_{di}} \frac{\rho_{high-k}(x)dx}{\epsilon_{di,high-k}} (t_{di} - x) \quad (2.66)$$

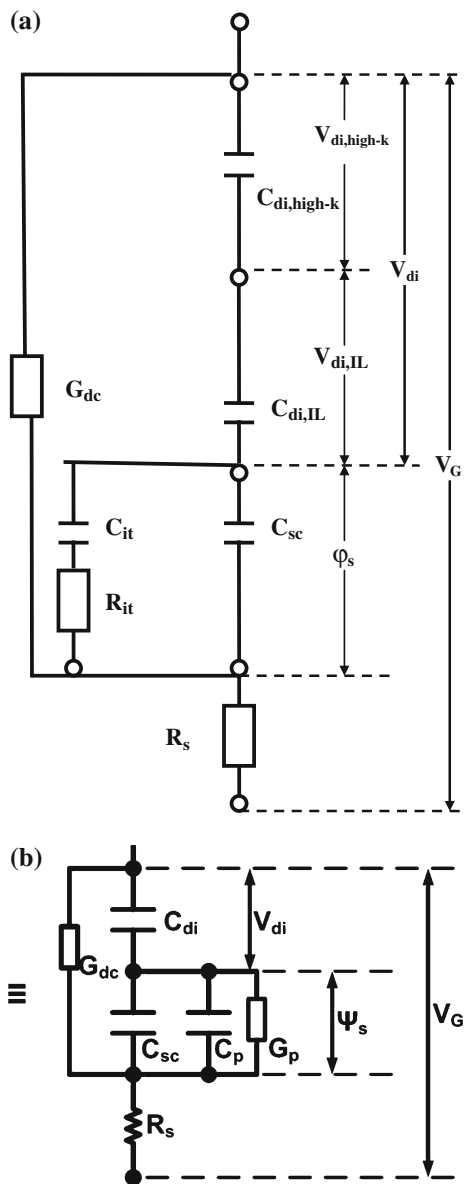
where $C_{di,high-k}$ is the dielectric capacitance, ρ_{high-k} is the volume charge density, and $\epsilon_{di,high-k}$ is the permittivity of this layer. The dielectric capacitance of a layer is its capacitance if it were charge-free, i.e. an ideal dielectric. This is the same as the plane parallel capacitance. The dielectric capacitances of the two layers of the gate stack may be expressed as:

$$C_{di,IL} = \frac{\epsilon_{di,IL}}{t_{di,IL}}; \quad C_{di,high-k} = \frac{\epsilon_{di,high-k}}{t_{di,high-k}}; \quad (2.67)$$

Just as the expressions for the potentials across the gate stack layers, a valid representation of the equivalent circuit of the gate stack is enormously more complicated than that in the case of the SiO_2 gate dielectric, as illustrated by Fig. 2.28. In this representation, R_s is the series resistance representing the entire device region outside of the silicon space charge layer and the gate stack, and is an important element in the case of large gate leakage currents. C_{sc} is the traditional silicon space charge layer capacitance, C_{it} is the traditional capacitance of the traps at the Si-SiO_x interface, R_{it} is the traditional Si-SiO_x interface trap resistance, and ϕ_s is the surface potential. G_{dc} represents the flow of the direct current through the gate stack [51]. Charging and discharging in traps in each of the two bulk layers of the gate stack and at two of the three interfaces has been represented by an $R_t C_t$ combination. It may be noted that there will be no charging or discharging of the traps on the metal surface (i.e. at the high-k/metal interface), as the metal Fermi level is bias invariant. An important issue in the case of charging/discharging of a trap inside the gate stack is where the free carrier is being supplied from, i.e. the silicon surface or the metal surface. If the free carrier exchange of the trap is with the silicon surface, then the $R_t C_t$ branch is connected to the silicon majority carrier band; if the free carrier exchange of the trap is with the metal surface, then the $R_t C_t$ branch is connected to the metal, cf. Fig. 2.28.

The circuit representation of Fig. 2.28 can be simplified, depending upon the bias and the small signal frequency. At a signal frequency too high for charging/discharging at any trap in the gate stack to follow, f_h , the circuit representation of the gate stack can be simplified to what is illustrated in Fig. 2.29. The circuit representation on the right side in Fig. 2.29b is close to the traditional

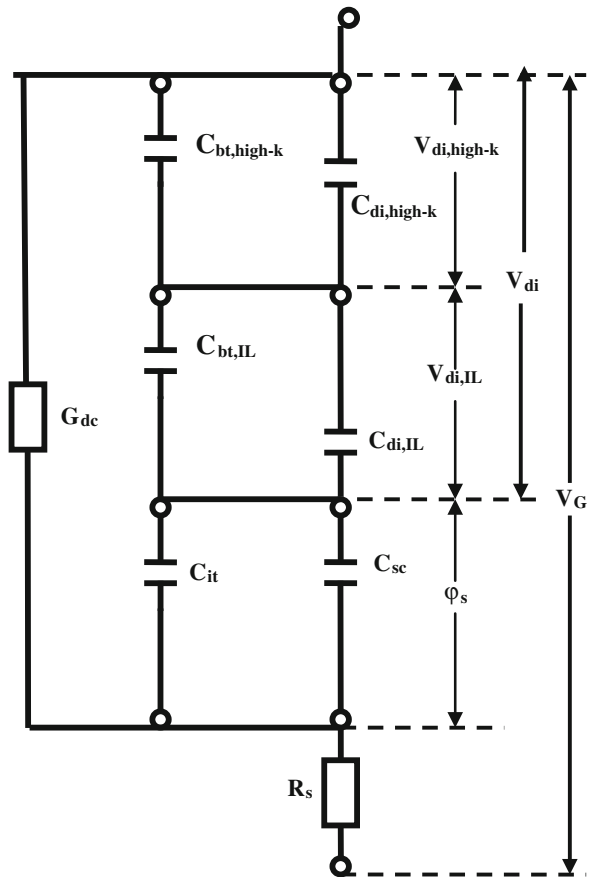
Fig. 2.29 **a** Reduction of the circuit representation in Fig. 2.28 at a high frequency, which no trap in the gate stack can follow.
b Transformation of **a**



representation, except that the total gate stack capacitance C_{di} is now a series sum of two plane-parallel capacitors:

$$\frac{1}{C_{di,hf}} = \frac{1}{C_{di,IL}} + \frac{1}{C_{di,high-k}} \quad (2.68)$$

Fig. 2.30 Reduction of the circuit representation in Fig. 2.28 at a low frequency, which all traps in the gate stack and the interface traps can follow



This is likely to be case, for all bias values, at the MOSFET clock frequency of a few GHz. At a signal frequency low enough for charging/discharging at every trap in the gate stack to follow, f_1 , the circuit representation of the gate stack can be simplified to what is illustrated in Fig. 2.30, in which case the total MOS capacitance would be given by:

$$\frac{1}{C_{lf}} = \frac{1}{C_{sc} + C_{it}} + \frac{1}{C_{di,IL} + C_{bt,IL}} + \frac{1}{C_{di,high-k} + C_{bt,high-k}} \quad (2.69)$$

2.8.1 Flat-Band Voltage Characteristics of High-k Gate Stack

SiO₂ Single Gate Dielectric.—As already outlined, the flat-band voltage V_{FB} is a very frequently used indicator of the MOS device quality; the use of V_{FB}

permeates most of this book. The flat-band voltage characteristics, in particular, its variation with the gate dielectric thickness or the EOT, is also very useful for parameter extraction. In the case of the SiO₂ single gate dielectric, the V_{FB} versus the t_{di} characteristic, as represented by (2.15) and reproduced below:

$$V_{FB} = -\frac{Q_{it,fb} + Q_F}{\epsilon_{di}} t_{di} - \phi_{MS} \quad (2.15)$$

has been used very effectively to extract the silicon-metal work function difference ϕ_{MS} and the interface charge density at flat-band ($Q_{it,fb} + Q_F$) [9, 10]. The $V_{FB}(t_{di})$ plots were found to be straight lines over the entire dielectric thickness range, the intercept of which with the $t_{di} = 0$ line yielded the work-function difference ϕ_{MS} [10] and the slope of which yielded the interface charge at flat-band [9]. Wafers with graded SiO₂ thickness was used in the study [9, 10], which perhaps was instrumental in producing high quality straight-line characteristics. Some points are worth noting in this context. Often, the interface charge density at flat-band is mistakenly taken as the fixed charge density Q_F ; this is erroneous, as for the device quality SiO₂ layers, $Q_{it,fb}$ and Q_F are of the same order of magnitude. The other point is that if Q_F and the interface traps are invariant of the silicon doping, which is generally assumed to be true, then, $Q_{it,fb}$ for n-Si will be different from $Q_{it,fb}$ for p-Si, i.e. $Q_{it,fb}$ for n-Si will be more negative than $Q_{it,fb}$ for p-Si. Experimental results clearly revealed the $V_{FB}(t_{di})$ straight line for n-Si to have a smaller slope than the same for p-Si [9, 10]. Although Q_F has never been determined, as there exists no technique to do so, the common wisdom considers the fixed charge density Q_F to be positive in the case of the SiO₂ gate dielectric [3]; this will support the above experimental result.

High-k Gate Stack.—In contrast to the experience with the SiO₂ gate dielectric, in the case of even the device quality gate stacks, the experimental flat-band voltage V_{FB} versus the EOT characteristic is non-linear, and it becomes increasingly non-linear as the value of EOT becomes small; for EOT < 1.0 nm, V_{FB} rolls off, leading to the emergence of the phrase: “ V_{FB} roll-off”. Making the assumptions used for obtaining (2.65) and (2.66), we may express the flat-band voltage for a high-k gate stack as, cf. (2.15), (2.65), and (2.66):

$$V_{FB} = -\phi_{MS} - \frac{Q_{it,fb} + Q_F}{\epsilon_{SiO_2}} EOT - \frac{\int_0^{t_{di,IL}} \rho_{IL}(t_{di,IL} - x) dx}{\epsilon_{di,IL}} - \frac{\int_0^{t_{di,IL}} \rho_{IL} dx}{\epsilon_{di,high-k}} t_{di,high-k} - \frac{\int_{t_{di,IL}}^{t_{di}} \rho_{high-k}(t_{di} - x) dx}{\epsilon_{di,high-k}} \quad (2.70)$$

Even a casual look at (2.70) will suggest that, unless many parameters in (2.70) are insignificant, any plot of V_{FB} versus any dielectric thickness will not yield a linear characteristic, which can be used to extract ϕ_{MS} or any of the gate stack charge densities. Unfortunately, in the relation of (2.70), the largest magnitudes have those charges, which induce the largest non-linearity, namely, ρ_{high-k} and ρ_{IL} .

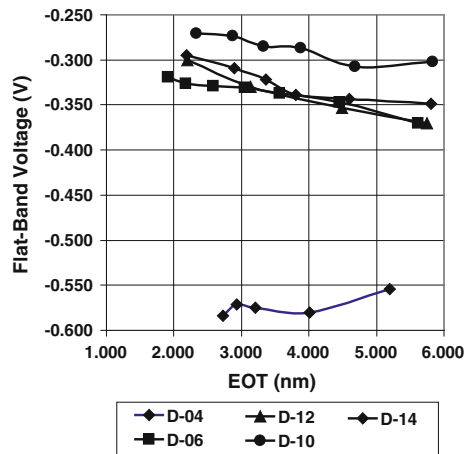
We may recall that in our formulation, both $\rho_{\text{high-k}}$ and ρ_{IL} include both the bulk and the interface trap charges. In general, the charge density as well as the permittivity increase moving from the Si/SiO₂ interface to the SiO₂ bulk to the SiO₂/high-k interface and finally to the high-k bulk. Consequently the potentials induced by the higher charges get partly neutralized by the corresponding higher permittivity, cf. (2.70).

In many experiments with high-k gate stacks, the high-k layer thickness is kept constant, while a graded SiO₂ intermediate layer is used to vary the IL thickness and thereby the EOT. In such a case, one would expect the $V_{\text{FB}}(\text{EOT})$ characteristic to be less non-linear, since the Si/IL interface is Si/SiO₂ and only the thickness of the SiO₂ layer is being varied. Even in such a situation, a linear $V_{\text{FB}}(\text{EOT})$ characteristic is not obtained, as is illustrated in Fig. 2.31. The wafers of Fig 2.31 had a graded SiO₂ layer (1–6 nm) grown in dry O₂ at 900 °C. Sample D-04 had no HfO₂ layer; samples D-06 and D-12 had 2 nm and samples D-10 and D-14 had 3 nm thick HfO₂ layer. Samples D-06 and D-10 had a post-deposition annealing (PDA) in O₂ at 500 °C for 1 min.

The results of Fig. 2.31 will be discussed in more detail later, see Sect. 2.10.3.2. Some relevant points worth mentioning here are:

1. As there are no data in the EOT range of 0–2 nm, the V_{FB} roll-off feature cannot be observed.
2. The inaccuracy in the flat-band voltage V_{FB} can be of the order of a few mV. This could have partly contributed to the scatter and the non-linearity of the $V_{\text{FB}}(\text{EOT})$ characteristics, as the change in V_{FB} is in the range of 10–20 mV/nm.
3. For the samples of Fig. 2.31, the SiO₂ trap density obtained from the conductance technique at about 0.25–0.30 eV above E_v was rather low and varied between 0.6 and $2.6 \times 10^{11} \text{ cm}^{-2} \text{ V}^{-1}$, depending upon the SiO₂ layer thickness, see Sect. 2.10.3.2. The trap density increased with decreasing SiO₂ layer

Fig. 2.31 Flat-band voltage versus EOT for p-Si/SiO₂/HfO₂/TaN MOS capacitors on different graded-SiO₂ wafers with different sets of PDA and HfO₂ thickness. The lower most curve belongs to wafer D-04 with no hafnia layer. The upper four curves belong to wafers with the hafnia layer [17]



thickness. As will be discussed in a later section, these traps may be located 0.4 nm inside the SiO₂ layer from the Si surface; this suggests that the traps at the Si surface and in the SiO₂ layer may depend upon the SiO₂ layer thickness. This would contribute to the non-linearity of the $V_{FB}(EOT)$ characteristics in Fig. 2.31.

5. Wafer D-04 has a positive gradient, while the other wafers have a negative gradient, suggesting that the flat-band interface charge is positive in the presence of the hafnia layer, while it is negative without.
6. The approximate flat-band interface charge density obtained from the slope of the $V_{FB}(EOT)$ characteristic in Fig. 2.31 was -2.1×10^{11} charges.cm⁻² for wafer D-04, 3.2×10^{11} charges.cm⁻² for wafer D-06, 4.1×10^{11} charges.cm⁻² for wafer D-12, 2.1×10^{11} charges.cm⁻² for wafer D-10, and 3.3×10^{11} charges.cm⁻² for wafer D-14. As the HfO₂ layer thickness was not varied, these charges at flat-band are likely to include only the flat-band charges at the Si-SiO₂ interface and in the SiO₂ layer, if we assume that the charges in the HfO₂ layer and at its two interfaces are not affected by the SiO₂ layer thickness, cf. Fig. 2.31.
7. There is a huge amount of charge in the hafnia layer and/or its two interfacial regions: one with the silica layer, and another with the TaN metal electrode; the difference between the flat-band voltage of wafer D-04 (no HfO₂ layer) and the other wafers is an indication of this charge, cf. Fig. 2.31. The total charge at flat-band in the HfO₂ layer and its two interfaces is net negative and its magnitude is roughly around $q \times 1.5 \times 10^{13}/\text{cm}^2$.

2.9 Impedance Characteristics of Leaky High-k MOS Structures

There are characteristic differences between the admittance characteristics of MOS devices with ultra-thin (say $EOT < 1.5$ nm) high-k gate stacks and those of MOS devices with non-leaky thicker (say $EOT > 6$ nm) gate dielectrics. The admittance characteristics undergo alterations as the gate dielectric thickness is reduced, say, from 6.0 to 1.5 nm. Some changes in the characteristics occur gradually, whereas some changes occur more drastically as a threshold thickness is crossed. One observes several unusual features in the impedance characteristics of ultrathin gate stacks on silicon channels, e.g. large accumulation and strong inversion regimes, flat-band voltage in a much less steep region, and narrow depletion and weak inversion regimes. In the case of ultrathin gate stacks on high mobility channels—such as Ge, GaAs, and InGaAs channels—several anomalous features, such as frequency dispersion of the accumulation and strong inversion capacitance, are observed, in addition to the salient features seen in the case of the silicon channels.

The classical electrical characterization tools were developed, for the MOS devices with the single SiO₂ gate dielectric, more than 3–4 decades ago in the

golden age of the MOS and the MOSFET research [1–4]. These tools are being employed for the leaky ultrathin high- k gate stacks without any significant change or adaptation. Many of these tools, such as the quasi-static C–V technique, simply do not function in the case of the leaky ultrathin gate stacks. Some other tools need to be modified to yield reliable values of the parameters. The electrical characterization tools need to be tuned and matched to the nature of the electrical characteristics. As there are features in the electrical characteristics, which are new, an opportunity and a scope exist for the development of new techniques for measurements and for analysis. Following is an analysis of the new features in the characteristics of the high- k gate stacks.

2.9.1 Si Channels

We will first analyze the admittance characteristics of the high- k gate stacks on the silicon substrate. The MOS characteristics on the high mobility substrates have always been a challenge to interpret and understand; these will be taken up in a later section.

2.9.1.1 Capacitance–Voltage (C–V) Characteristics

Figure 2.32—normalized capacitance, C/C_{di} , versus voltage V —and Fig. 2.15—the corresponding surface potential ϕ_s versus voltage V —illustrate some of the salient features of the characteristics of the MOS devices with ultra-thin high- k gate dielectrics. Figures 2.32 and 2.15 represent four carefully chosen devices in MOSFET and MOS capacitor configuration with different ultrathin ($EOT = 0.46$ – 1.94 nm) high- k materials (HfO_2 , $HfAl_2O_5$, La_2O_3 , $HfSiON$) as gate dielectrics, different high- k deposition methods (ALD, e-gun evaporation, oxidation), and gate electrode materials (poly-Si, Ti, Al, TiN), offering a wide variation in band offsets ($\phi_b = 2.00$ – 4.19 eV), effective tunneling mass ($m^* = 0.22$ – 0.46 m), and dielectric constant ($k = 14$ – 33); the measured C–V data of these devices were taken from the literature [56–59] as indicated in the caption of Fig. 2.15. The C–V characteristic of the MOS capacitor with the SiO_2 gate dielectric has been included for the sake of comparison. Indicated, in Fig. 2.32, are the flat-band voltage, and also the voltage for the onset of strong inversion; these voltages were obtained from the $\phi_s(V)$ characteristics of Fig. 2.15, which was extracted by the integration of the measured C–V characteristic, in accordance with the Berglund relation [60], cf. Sect. 2.10.2.

One salient feature that one can observe in Fig. 2.32 is that, most of the C–V characteristic is in the accumulation regime and the strong inversion regime. For example, in the case of the MOS transistor with the HfO_2 – Al_2O_3 gate dielectric, the accumulation and the strong inversion regime cover a voltage range of nearly 3.62 V, while the depletion and the weak inversion regimes cover only a fraction

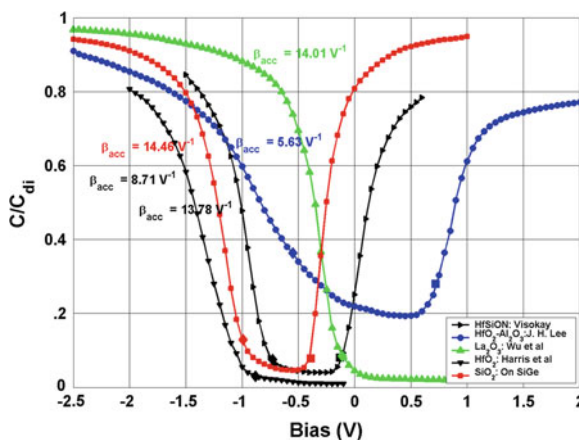


Fig. 2.32 High frequency normalized capacitance (C/C_{di}) versus bias for four MOS capacitors or transistors on p-type silicon with four different high-k gate stacks as indicated in the legend, and for the control sample with the SiO_2 gate dielectric: HfSiON (right faced triangle, black), $\text{HfO}_2\text{-Al}_2\text{O}_3$ (circle, blue), La_2O_3 (triangle, green), HfO_2 (inverted triangle, black), and SiO_2 (square, red) with EOT values of 2.0, 1.7, 1.4, 0.5, and 3.9 nm respectively. It maybe noted that the diamond marker indicates the flat-band point, while the square marker indicates the onset of strong inversion. The C–V data for four of the characteristics were taken from the literature: HfSiON [56], $\text{HfO}_2\text{-Al}_2\text{O}_3$ [57], La_2O_3 [58], HfO_2 [59]

of it, namely, about 1.38 V. Secondly, a lower slope in the falling or the rising parts of the C–V curve does not reflect or represent a higher density of interface states in the silicon band-gap, as it used to be the case for the classical MOS devices. It is not clear yet what this slope represents or reflects; a lower slope of C–V characteristic in accumulation or in strong inversion may reflect a higher density of states inside the conduction or the valence band and/or a low direct tunneling current index [22]. Thirdly, certain traditional MOS parameter extraction techniques, if applied, would lead to less reliable results. For example, as the quasi-static C–V technique is not available, often, one finds recourse to the Terman technique [61] in the current literature for extracting the interface state density D_{it} . For the ultra-thin gate dielectrics, it can be shown that the Terman technique is very unreliable, even when D_{it} is high, because of a small potential across the gate stack and a strong and uncertain doping density profile.

In the case of the classical MOS devices, the C–V characteristics were dominated by the depletion and the weak inversion regimes; these regimes yielded almost all the information that could be extracted. The situation is almost reversed in the case of the high-k gate dielectrics of Fig. 2.32, with the accumulation and the strong inversion regimes dominating the C–V characteristics and offering a huge window, which permits a deep look into the accumulation and the strong inversion layers. This opens up the scope of extracting significantly more information (than allowed before) from these two regimes and also for the development of new parameter extraction techniques to make use of this opportunity.

Significance of the parameter—accumulation surface potential index β_{acc} , values of which have been indicated in Fig. 2.32, will be discussed in Sect. 2.10.2. This index was found to vary with the high-k material and its processing and may represent the quality of the high-k gate stack [22].

Noteworthy in Fig. 2.15 are the large surface potential ranges in the accumulation and the strong inversion regimes, particularly when the accumulation surface potential index is small, e.g. see the enormous accumulation surface potential range of 0.62 V for the MOS transistor with the HfO_2 high-k layer. It may be seen that in the depletion and the weak inversion regimes, φ_s is linear with V, with a slope very close to unity, except for the MOSFET with the $\text{HfO}_2\text{--Al}_2\text{O}_3$ gate dielectric. The parallel displacement of one $\varphi_s(\text{V})$ from the other reflects the difference in the flat-band voltage. As already discussed in Sect. 2.6.1, the non-saturating surface potential in both strong inversion and accumulation regimes is a strong departure from the classical MOS behavior, cf. Fig. 2.15.

2.9.1.2 Conductance–Voltage (G–V) Characteristics

Figure 2.33 presents the capacitance–voltage (C–V) characteristics of a p-Si/ SiO_2 / HfO_2 /TaN MOS capacitor, measured at 10 and 100 kHz, respectively, and Fig. 2.34 represents the corresponding G–V characteristics. The physical thickness of the HfO_2 layer was 2 nm, and the EOT of the gate stack was 1.9 nm. In Fig. 2.33, the slightly higher plot in the accumulation regime, represents the lower frequency (i.e. 10 kHz), while in Fig. 2.34, the higher plot represents the higher frequency (i.e. 100 kHz). Indicated in both Figs. 2.33 and 2.34 are the flat-band point V_{FB} and the accumulation and the depletion regimes; the surface potential was extracted from the Berglund integral [60]. It may be noted that the conductance observed in Fig. 2.34 in accumulation and also the slight frequency dispersion of the accumulation capacitance observed in Fig. 2.33 are due to the series resistance. The observed right peak in conductance, in the weak inversion regime,

Fig. 2.33 Capacitance–voltage (C–V) characteristics of a p-Si/ SiO_2 / HfO_2 /TaN MOS capacitor, measured at 10 and 100 kHz, respectively [72]

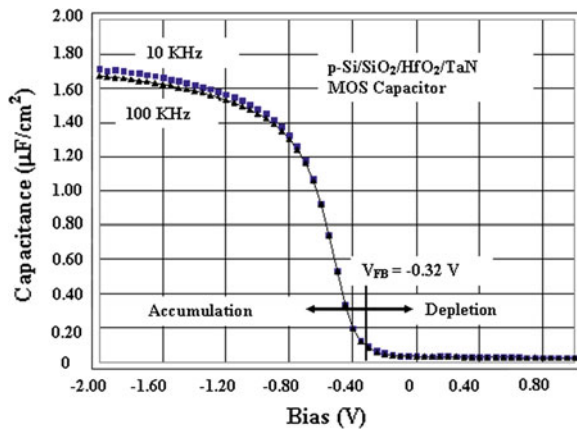
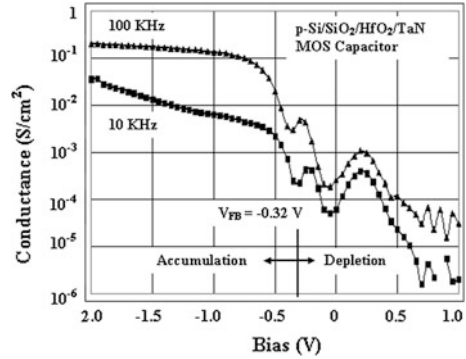


Fig. 2.34 Conductance–voltage (G–V) characteristics of a p-Si/SiO₂/HfO₂/TaN MOS capacitor (the same as in Fig. 2.33), measured at 10 and 100 kHz, respectively [72]



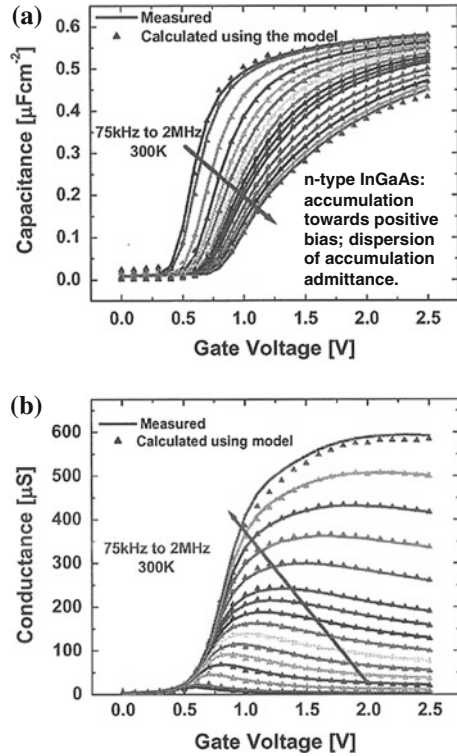
is due to the partial (lossy) response of the inversion layer. In the depletion regime, the conductance peak (the left peak) observed is a manifestation of the loss involved in the capture of holes by the traps at some location x_t in the gate stack. These points will be discussed in more detail in Sect. 2.10. The conductance–voltage (G–V) characteristics of ultrathin leaky gate stacks are complicated by the effect of the series resistance in the accumulation regime and the lossy response of the inversion layer in the weak inversion regime. These features are generally not observed in the case of device grade non-leaky classical gate dielectrics.

2.9.2 Ge and III–V Compound Semiconductor Channels

As indicated already, the admittance–voltage–frequency characteristics of ultrathin high- k gate stacks on high mobility substrates—such as Ge, GaAs, InGaAs—exhibit several unusual features [62, 63]. One of these is the strong frequency dispersion of the accumulation and also the strong inversion capacitance; it may be added that this dispersion has been observed for gate dielectrics—such as 1-octadecene monolayers—on silicon substrates [64]. The capacitance–voltage characteristics of Fig. 2.35 illustrate the huge frequency dispersion of the accumulation as well as the depletion capacitance. The latter is most likely caused by the very high density of the interface traps—exceeding $10^{13} \text{ cm}^{-2} \text{ V}^{-1}$. Several models [62, 63] have been proposed to explain the frequency dispersion of the accumulation and the strong inversion capacitance; however, this phenomenon is not yet adequately explained. The following is a possible interpretation of this unusual feature.

In the case of the classical gate dielectrics and MOS devices, the accumulation or the strong inversion capacitance approached the capacitance of the gate dielectric, which was assumed to be frequency-independent, as the capacitance of an ideal dielectric should be. The leaky high- k gate stack is a strong departure from an ideal gate dielectric—it has a very high density of traps and as the gate stack is ultrathin and leaky, many of the traps inside the gate stack can exchange

Fig. 2.35 C–V and G–V Characteristics of $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}/\text{LaAlO}_3/\text{TaN}$ gate stacks measured at different frequencies in the range of 75 kHz and 2 MHz, illustrating frequency dispersion of the accumulation capacitance and conductance. Adapted from [62]



electrons or holes with the semiconductor or the metal surface, thereby contributing and adding trap capacitance to the dielectric capacitance of the gate stack, cf. Sect. 2.8, Figs. 2.28, 2.29 and 2.30. When the ac signal frequency is high enough, the total capacitance of the gate stack reduces to its high frequency value, see Fig. 2.29 and (2.68). As the frequency is reduced, the gate stack capacitance increases because its dielectric capacitance is augmented by the charging capacitance of the traps, see Fig. 2.30 and (2.69). The frequency dispersion of the accumulation conductance, cf. Fig. 2.33, may be caused partly by the charging or discharging loss at the gate stack traps; other sources of this dispersion may be the series resistance and the leakage current.

2.10 Parameter Extraction Techniques

As already outlined, extraction of high- k gate stack parameters is confronted with several difficult roadblocks: (1) the high- k gate stack is a much more complicated system than the SiO_2 gate dielectric; (2) many of the classical MOS parameter extraction techniques are disabled by the nature of the leaky ultrathin high- k gate

stack; (3) no new parameter extraction technique has so far been successfully developed for the high- k gate stack; (4) the frequency range of measurements should extend up to GHz. Ideally, we should have techniques which cover very low to the microwave range: (1) static (10^{-5} to 10^{-2} Hz); (2) quasi-static (10^{-3} to 10^{-1} Hz); (3) LF (Low Frequency – 10 Hz to 100 kHz); (4) HF (High Frequency – 1 to 50 MHz); (5) Ultra High Frequency (UHF – 100 to 900 MHz); (6) Microwave (1 to 20 GHz). We need the static and the quasi-static frequencies to access traps deep inside the high- k gate stack and also to get full response of the inversion layer. We need the HF, UHF, and microwave frequencies to examine the dependence of the dielectric constant on frequency and also to be able to measure the leaky capacitance more reliably. We need the LF, HF, and the UHF frequencies to obtain the full range of the G - V - f characteristics and the full range of the G_p/ω peaks. Because of the gate leakage current, measurements at static, quasi-static, and even most of the LF ranges are not possible. These acute constraints make parameter extraction a difficult exercise in the case of the high- k gate stacks.

2.10.1 Determination of the High- k Gate Stack Capacitance C_{di}

The gate stack capacitance, C_{di} , is a crucial parameter, as it is directly related to the drive current, the trans-conductance, the channel conductance, and the threshold voltage. Furthermore, it reflects somewhat the physical thickness of the gate stack, which is the most important parameter in deciding all phenomena related to quantum-mechanical tunneling (wave-function penetration, carrier confinement, tunneling current, wave-function mixing). Therefore, accurate determination of C_{di} is a crucial exercise. This exercise used to be rather simple in the case of thick gate dielectrics, as the MIS capacitance in very strong accumulation, C_{acc} , saturated nicely to the gate dielectric capacitance C_{di} . In the case of ultrathin, say EOT <3 nm, gate dielectrics, extraction of C_{di} is beset with several problems. The accumulation capacitance does not saturate to C_{di} , because C_{di} is no longer insignificant in comparison to the space charge capacitance in accumulation, and also due to the carrier confinement in the accumulation layer. An ever-increasing gate dielectric leakage current density and the parasitic impedance of the substrate, also make the accumulation capacitance C_{acc} differ from its ideal value. Other factors which may cause C_{acc} to deviate significantly from the gate stack capacitance C_{di} include: (a) capacitance of traps in the gate stack; and (b) low density of states in the conduction band, in the case of compound semiconductors.

Several capacitance techniques [65–69] exist, which were originally developed to extract the capacitance of SiO_2 gate dielectrics; three of these extract the dielectric capacitance directly, while the fourth involves modelling and curve-fitting. These techniques operate under various assumptions, all of which are

invalid in the case of the high-k gate dielectrics. Three of the main differences between SiO₂ and high-k gate dielectrics, which make the above assumptions invalid, relate to the Si/IL interface trap density, D_{it} , which is higher in the latter case, the charge densities in the high-k gate stack, which are orders of magnitude higher in the latter case, and the conductance or the valence band offsets, $\phi_{b,c}$ or $\phi_{b,c}^*$, which are much lower in the latter case. In these techniques, the interface trap charge, Q_{it} , is neglected in accumulation and around flat-band. Even, in the case of the SiO₂ gate dielectric, D_{it} can be much higher near the band edges; in the case of the high-k gate dielectrics, D_{it} and the corresponding Q_{it} can be too large in accumulation to neglect. The interface trap capacitance, C_{it} , is neglected in the 100 kHz or the 1 MHz (taken as high frequency) accumulation capacitance. It can be easily shown that the interface traps are likely to follow the 1 MHz ac signal in accumulation. The charges in the high-k gate stack are simply too large to ignore. The above-mentioned techniques briefly are:

1. *McNutt and Sah Technique* [65]—One plots $|dC/dV|^{1/2}$ versus C in the accumulation regime. If a linear fit is obtained, then its intercept with the C axis yields C_{di} .
2. *Maserjian Technique* [66, 67]—One plots $|dC^{-2}/dV|^{1/4}$ versus $1/C$ in the accumulation regime. If a linear fit is obtained, then its intercept with the C^{-1} axis yields $1/C_{di}$.
3. *Ricco Technique* [68]—In this technique, a factor, F_{Ricco} is calculated around the flat-band condition. The flat-band voltage, V_{FB} is determined from the condition that for $V = V_{FB}$, $F_{Ricco} = 0$. Subsequently, the gate dielectric capacitance is calculated from the MIS flat-band capacitance and the flat-band space-charge capacitance, calculated using an approximate relation, obtained by Ricco et al. [68].
4. *Curve-Fitting Technique* [69]—In this technique, a calculated capacitance–voltage, C – V , characteristic is matched to the measured 100 kHz or 1 MHz C – V , adjusting a large number of fitting parameters, to yield the values of the physical (i.e. the fitting) parameters. There can be justifiable concerns regarding all the assumptions of this technique, as well as the technique itself of adjusting many fitting parameters, the number of which will be higher, and their independent measurements more difficult in the case of the high-k gate dielectrics. Particularly, difficult is the determination of the physical parameters (dielectric constants and thicknesses) of gate dielectric stacks, where significant interlayer diffusion and interlayer chemical reactions have taken place, by design or otherwise. Assumption of an infinite potential barrier at the interface is questionable, as the conductance and valence band offsets are in many cases around 2 eV or even less, cf. Appendix V. At a signal frequency of 100 kHz or 1 MHz, there may be significant contribution to the accumulation gate stack capacitance from the traps inside the gate stack, as has been discussed in Sect. 2.8, which the curve-fitting model does not take into account.

2.10.1.1 A New Direct Capacitance Extraction Technique

More recently, a capacitance technique has been proposed to better address the current problems; this technique yields the gate stack capacitance directly, and makes none of the above assumptions. In this technique, two simple options are available for the determination of the capacitance of a leaky ultrathin gate dielectric, using mathematical relations in closed form [21]. These relations [21], presented below, are valid in strong accumulation, if the corresponding parallel capacitance, $C_{p,acc}$, which is the sum of the space charge capacitance, $C_{sc,acc}$, and the interface trap capacitance, $C_{it,acc}$, is an exponential function of the surface potential φ_s , i.e. $C_{p,acc} \propto \exp(\beta_{acc}\varphi_s)$, which was confirmed convincingly by experimental results on a variety of MOS structures with high-k gate stacks, cf. Fig. 2.36.

$$\left| \frac{1}{C} \frac{dC}{dV} \right|^{\frac{1}{2}} = \frac{|\beta_{acc}|^{\frac{1}{2}}}{C_{di}} (C_{di} - C) = \frac{-\sqrt{|\beta_{acc}|}}{C_{di}} C + \sqrt{|\beta_{acc}|}. \quad (2.71)$$

$$\frac{1}{C} = \frac{1}{C_{di}} + \left| \frac{1}{2\beta_{acc}} \frac{d}{dV} \frac{1}{C^2} \right|^{\frac{1}{2}} \Rightarrow \left| \frac{dC^{-2}}{dV} \right|^{\frac{1}{2}} = \frac{\sqrt{2|\beta_{acc}|}}{C} - \frac{\sqrt{2|\beta_{acc}|}}{C_{di}}. \quad (2.72)$$

It can be seen from (2.71) that, a plot of $|C^{-1}dC/dV|^{1/2}$ versus C , in the accumulation regime, should result in a straight line, whose x-intercept would yield the gate dielectric capacitance C_{di} , whose y-intercept would yield $|\beta_{acc}|^{1/2}$, and whose slope (dy/dx) would yield $(-|\beta_{acc}|^{1/2}/C_{di})$. Linear plots (cf. Fig. 2.37) were obtained for all the high-k gate stacks presented in Table 2.1, from which values of C_{di} and β_{acc} were obtained from the x- and y-intercepts, respectively, cf. Table 2.1. Similarly, a plot (cf. Fig. 2.38) of $|dC^{-2}/dV|^{1/2}$ versus C^{-1} , in the accumulation regime, also resulted in a straight line, for the wide variety of gate stacks of Table 2.1. According to (2.72), the x-intercept of this line would yield C_{di}^{-1} , the y-intercept would yield $(-2|\beta_{acc}|^{1/2}/C_{di})$, and whose slope would yield $(2|\beta_{acc}|)^{1/2}$. Values of C_{di} and β_{acc} were obtained from the x-intercept and the

Fig. 2.36 Experimental parallel capacitance, in the strong accumulation regime, $C_{p,acc} = C_{sc,acc} + C_{it,acc}$, versus the surface potential, φ_s , for five MOS devices (on p-type silicon) containing different high-k gate stacks, cf. Table 2.1 [22]

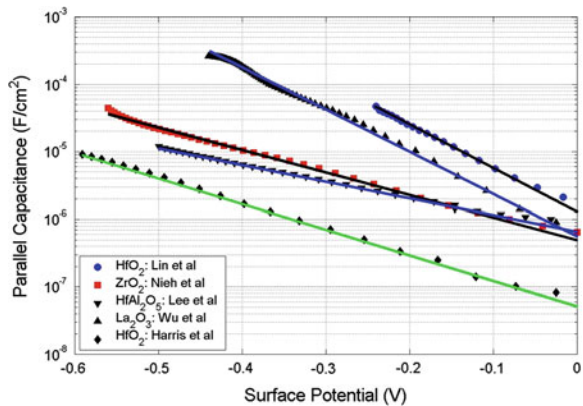
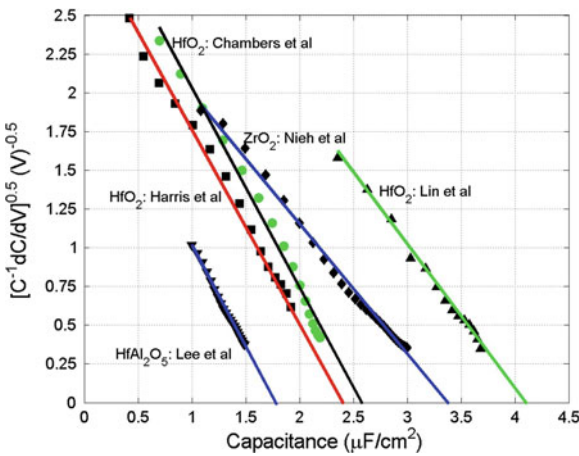


Fig. 2.37 $|C^{-1}dC/dV|^{1/2}$ versus capacitance C , in strong accumulation, for different high- k gate stacks, cf. Table 2.1. The intercept, of the linear fit to the data points, with the x-axis, yielded experimental values of C_{di} and its slope yielded experimental values of β_{acc} [21]



slope, respectively, of the straight lines in Fig. 2.38, cf. Table 2.1. It may be noted that experimental values of C_{di} or β_{acc} may be obtained from the slope of Fig. 2.37 or the y-intercept of Fig. 2.38; however, these may have lower accuracy.

Unsatisfactory results were obtained from the application of three of the existing direct capacitance techniques [65–68] to the high- k gate stacks. The McNutt and the Maserjian plots were very non-linear, thereby suggesting invalidity of the assumptions under which the approximate mathematical relations were derived, in the cases of high- k gate stacks. Different linear fits could be made to different parts of the plots, i.e. in different parts of the accumulation regime, but, this led to ambiguity and improbable and multiple values for the dielectric capacitance. Figure 2.39a and b illustrate some of the deficiencies and present some comparison of the efficacy of the McNutt and the Maserjian Techniques with the technique proposed by Kar [21].

Table 2.1 Reference [21]: experimental values of the gate stack capacitance C_{di} and the exponential index of the surface potential φ_s of MOS structures with a wide variety of high- k gate stacks, with different composition, and/or band offsets, and/or deposition conditions, fabricated by various research groups [57–59, 82–84]

Substrate/High-K/Gate electrode	From Fig. 2.37		From Fig. 2.38		From Fig. 2.36
	C_{di} ($\mu\text{F}/\text{cm}^2$)	β_{acc} (V^{-1})	C_{di} ($\mu\text{F}/\text{cm}^2$)	β_{acc} (V^{-1})	β_{acc} (V^{-1})
p-Si/HfAl ₂ O ₅ /poly-Si [57]	1.76	−5.54	1.77	−5.37	−5.63
p-Si/ZrO ₂ /TaN [82]	3.43	−7.60	3.41	−7.73	−7.81
p-Si/HfO ₂ /Al [59]	4.08	−14.47	4.09	−14.47	−14.49
n-Si/HfO ₂ /TaN [83]	2.91	8.50	2.98	7.75	9.06
p-Si/La ₂ O ₃ /Al [58]	7.36	−13.38	7.37	−13.14	−14.01
p-Si/HfO ₂ /Ti [84]	2.47	−8.76	2.48	−8.66	−8.71

Fig. 2.38 $|dC^{-2}/dV|^{1/2}$ versus inverse capacitance C^{-1} , in strong accumulation, for different high-k gate stacks, cf. Table 2.1. The intercept, of the linear fit to the data points, with the x-axis, yielded experimental values of C_{di}^{-1} and its slope yielded experimental values of β_{acc} [21]

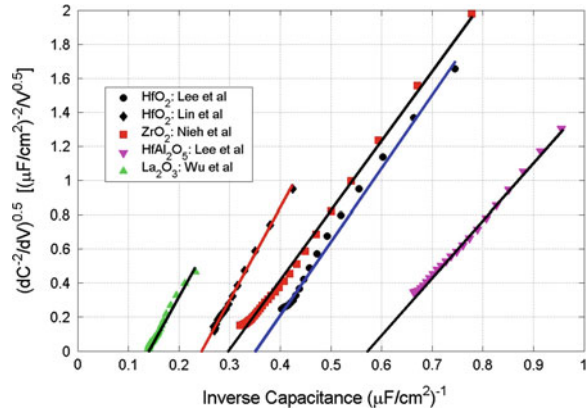


Figure 2.39a illustrates the comparison of the quality of the plots and the attendant results between the Kar technique and the McNutt and Sah technique in the case of a lanthanum oxide gate dielectric with an EOT of 0.48 nm. (p-Si/La₂O₃/Al structure by oxidation of lanthanum on silicon; 3.3 nm La₂O₃; gate area of 1×10^{-4} cm²). It can be seen in Fig. 2.39a that all the data points from the Kar technique, i.e. almost the entire accumulation regime, fall on a straight line. The McNutt data points lie on a very non-linear curve, which is likely to reflect the neglect of trap charges and states. The value of the dielectric capacitance, obtained from the intercept of the linear fit to the data points of the Kar technique is 722.22 pF. The highest accumulation capacitance measured was 708.61 pF at -3.50 V. A value of 0.48 nm is obtained for Capacitive Equivalent Thickness (CET), corresponding to a C_{di} of 722.22 pF. The value of β_{acc} obtained from the slope of the linear fit is 15.61 V⁻¹. If one makes a linear fit to the McNutt data points in very strong accumulation, i.e. to only a part of the curve, one obtains a value of 734.21 pF for C_{di} .

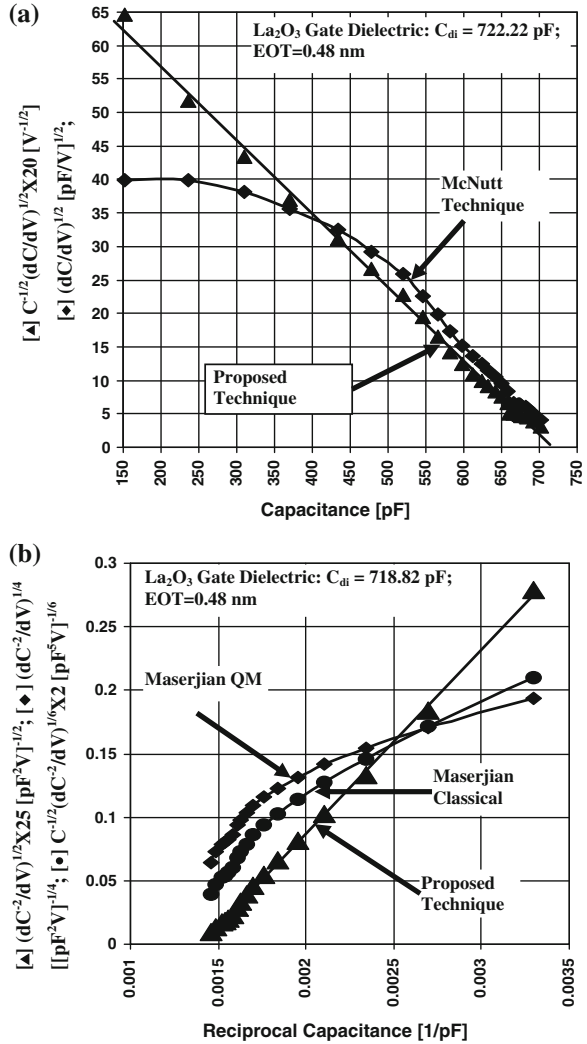
Figure 2.39b illustrates the quality of the plots and the attendant results from the Kar technique in comparison to the Maserjian technique for the same sample as in Fig. 2.39a. It can be seen that the data points from the Kar technique fit a straight line even better, over the entire accumulation regime, than in Fig. 2.39a. The value of the dielectric capacitance obtained from the intercept of the linear fit to the data points of the Kar technique is 718.82 pF, while a value of 16.93 V⁻¹ is obtained for β_{acc} from its slope. The Maserjian data points lie on a very non-linear curve. Absurd values of C_{di} result from linear fits to parts of the curve in accumulation or moderate accumulation. The best values are obtained from linear fits to the curve in very strong accumulation, which are 945.63 pF from the quantum-mechanical Maserjian technique, and 845.67 pF from the classical Maserjian technique.

The Kar gate stack capacitance extraction technique [21] has since been applied to a very large number of high-k MOS structures [15]; this technique has worked well in all the cases applied, irrespective of the gate stack composition, the

Fig. 2.39 (a, top)

Comparison of the quality of plots obtained from the Kar (proposed) technique versus the McNutt technique for a p-Si/La₂O₃/Al structure (3.3 nm La₂O₃). The intercept from the Kar technique yields a C_{di} of 722.22 pF. The slope yielded a value of 15.61 V⁻¹ for the exponential constant β_{acc} .

(b, bottom) Comparison of the quality of plots obtained from the Kar (proposed) technique versus the classical and the quantum-mechanical (QM) Maserjian techniques for the same gate dielectric as in Fig. 2.39a. The intercept from the proposed technique yields a C_{di} of 718.82 pF. The slope yielded a value of 16.93 V⁻¹ for the exponential constant β_{acc} [21]



deposition/fabrication conditions, and the value of EOT, consistently producing linear plots over the entire accumulation region, in accordance with (2.71) or (2.72); also, the extracted accumulation capacitance was without any exception observed to be an exponential function of the surface potential.

2.10.1.2 The Curve-Fitting Capacitance Extraction Technique

Perhaps, the most frequently applied and the most popular gate stack capacitance extraction technique, currently, is the curve-fitting technique [40, 69] or some variation of the same. It is not clear what the reason for this popularity and what

the attraction of this technique could be. At the present time, no option exists to ascertain the reliability or the veracity of the values of C_{di} and EOT obtained from the curve-fitting technique, or for that matter, from any other C_{di} extraction technique. Physical thicknesses of the different layers of the gate stack could be obtained from the cross-sectional transmission electron micro-graphs, perhaps, not too inaccurately. Also, a value of the EOT can be calculated from the extracted gate stack capacitance density C_{di} , but the connection of the EOT to the physical thicknesses of the gate stack, cf. (2.73), is not possible, as it is beyond our reach at the moment to determine the values of the dielectric constant of the individual gate stack layers.

$$EOT = \frac{\varepsilon_{SiO_2}}{C_{di}} = t_{di,IL} \frac{k_{SiO_2}}{k_{IL}} + t_{di,HfO_2} \frac{k_{SiO_2}}{k_{HfO_2}} + t_{di,cap} \frac{k_{SiO_2}}{k_{cap}} \quad (2.73)$$

In (2.73), ε_{SiO_2} represents the electrical permittivity of the SiO_2 , the respective t 's represents the physical thickness, and the respective k 's represents the dielectric constant of the different gate stack layers: IL, HfO_2 , and the cap layer, if a cap layer is present. Therefore, as there is no independent verification of the accuracy of the generic curve-fitting technique, it could not be that the curve-fitting technique is popular, because it has been proved to be a reliable technique.

One possibility for the popularity of the generic curve-fitting technique could be that it is easy to foresee that the technique could yield a significantly lower value for the EOT than what the C–V characteristic would realistically suggest and what the real value could be. The curve-fitting technique may be overestimating the carrier confinement effects, overestimating the effect of the parasitic resistance, and may be overestimating the effect of the gate leakage current, on the gate stack capacitance. All of these three factors, if present, make the measured accumulation capacitance lower than the gate stack capacitance.

1. As already analyzed, carrier confinement is significantly diluted by a moderate band offset (in place of an infinite barrier, generally assumed in the curve-fitting technique), is significantly diluted by significant wave-function penetration in deep accumulation and deep inversion, is significantly diluted by the tunneling current, and is significantly diluted by the mixing with the metal wave function.
2. A series resistance is a very poor representation of the parasitic impedance of the passive elements of the CMOSFET, which are vast in size and in number. Perhaps, more importantly, it is impossible to measure it. Even in a non-leaky MOS structure (say, with an EOT of 4.0 nm) with a SiO_2 gate dielectric, one does not measure a frequency-independent value of the series resistance in deep accumulation. Resultantly, the modelling of the series resistance effect in the curve-fitting technique is rather baseless.
3. Likewise the carrier confinement effect, both the series resistance and the gate stack leakage current make the measured accumulation capacitance to be lower than C_{di} in the deep accumulation regime. Similar is the effect of inductance at high frequencies (1 MHz). The modelling of the leakage current effect in the curve-fitting technique is too empirical and has no sound basis.

4. There has been no experimental verification of the models of any of the above three effects on the C–V characteristic, as it is not possible to separate these effects from one another or from that of the parasitic inductance. A drooping C–V characteristic in the deep accumulation regime could be an indication of the series resistance, and/or, the leakage current, and/or the parasitic inductance effect.
5. As already discussed, the deep accumulation capacitance may include significant contribution from the charging capacitance due to trapping or de-trapping inside the gate stack. This also will give a higher gate stack dielectric capacitance density and a lower EOT than what the reality is. It is important to note that, at the operating frequency (GHz range), there will be no trapping or de-trapping inside the gate stack; so, what will determine the drain current, transconductance, switching time, etc., will be the dielectric capacitance of the gate stack, and without any contribution from the gate stack trap capacitance.

2.10.2 Extraction of the Surface Potential ϕ_s

The surface potential (same as the semiconductor band bending or the interface potential) ϕ_s is one of the most useful parameters in analyzing the MOSFET function.

1. Knowledge of the surface potential enables us to know whether the MOSFET is operating in the accumulation, in the depletion, in the weak inversion, or in the strong inversion regime.
2. Knowledge of the surface potential is necessary for determining the interface trap energy.
3. Knowledge of the surface potential is indispensable for using the MOS conductance technique, and for determining the interface state time constant and the interface state capture cross-section.
4. An accurate value of the surface potential enables us to calculate the free carrier density at the semiconductor surface, and hence the free carrier density at the site of the gate stack trap.
5. Knowledge of the surface potential enables us to know the profile of the carrier confinement potential well.

There is only one way for extracting the surface potential accurately, i.e. by an integration of the equilibrium (or the low frequency) capacitance–voltage (C–V) characteristic. An equilibrium C–V is obtained when both the traps and the minority carriers can follow the applied small signal. As the minority carriers contribute to the capacitance only in strong and weak inversion, the signal needs to be followed by the minority carriers only in strong and weak inversion, but, by the traps for all the bias values. A typical frequency for measuring the equilibrium C–V is of the order of 10^{-3} Hz. Although, sinusoidal frequencies as low as 10^{-5} Hz

are available, small signal C–V measurement is difficult below 100 Hz, because of noise and also because of signal source instability. Therefore the equilibrium C–V is generally obtained from the quasi-static or the static measurement. In the quasi-static measurement, a ramp voltage, $V = at$, where a is a constant, is applied and the charging current is measured, which yields the quasi-static capacitance:

$$C = \frac{\partial Q}{\partial V} = \frac{\partial Q}{\partial t} \frac{\partial t}{\partial V} = \frac{I_{\text{charging}}}{\partial V / \partial t} = \frac{I_{\text{charging}}}{a}$$

A typical ramp rate is 1 mV/s. For the measurement samples (i.e. the test structures), the quasi-static charging current is typically in the pA range. The gate stack leakage current, if significant, adds to the charging current; unless the gate leakage current is significantly smaller than a pA, the quasi-static capacitance measurement becomes faulty. In the static C–V measurement, a voltage step (say, 10 mV) is applied; the corresponding incremental charge is measured after a time delay (say, 100 s), thereby yielding the static capacitance. For a reliable quasi-static or a reliable static measurement, it is essential that the gate leakage current is insignificant. This means that both the quasi-static and the static C–V measurements are not possible in the case of the ultrathin, leaky gate stacks. For the leaky gate stacks, the equilibrium C–V can be obtained only for the accumulation regime, for which the only option that is available is to measure the C–V at 1 kHz or so.

As the minority carrier contribution to the accumulation capacitance can be easily ignored, only the response of the majority carriers and the interface traps to the applied small signal, will decide the equilibrium frequency. The majority carrier response time in the silicon substrate is of the order of a ps or less; so, the majority carriers would have absolutely no problem with a 1 kHz or even a 100 kHz signal. The interface trap time-constant for hole capture may be expressed as:

$$\tau_h = \frac{1}{v_h \sigma_h p_s} \quad (2.74)$$

Assuming a hole capture cross-section of 10^{-15} cm^2 , and a hole density of 10^{16} cm^{-3} at the silicon surface, the interface trap time-constant estimates to 10 ns. For the interface traps at the Fermi level to follow the signal, the condition $\omega (=2\pi f) \ll (\tau_h)^{-1}$ has to be fulfilled; this is the case for $f = 100 \text{ kHz}$. Hence, the 100 kHz or a lower signal frequency can be considered to be an equilibrium frequency in accumulation.

The equilibrium C–V is to be integrated to obtain the surface potential ϕ_s , according to the relation [60]:

$$\phi_s - \phi_{s,0} = \int_0^V \left(1 - \frac{C}{C_{di}} \right) dV \quad (2.75)$$

where $\varphi_{s,0}$ is the surface potential at zero bias. The relation (2.75) can be derived from the circuit representation of Fig. 2.4b and the incremental voltage division defined in Fig. 2.3a:

$$\frac{d\varphi_s}{dV} = \frac{dV - dV_{di}}{dV} = 1 - \frac{dV_{di}}{dV} = 1 - \frac{dQ_M dV_{di}}{dQ_M dV} = 1 - \frac{C}{C_{di}} \quad (2.76)$$

It is apparent from (2.75) and (2.76) that the area above the C–V curve represents the surface potential φ_s while the area below represents the gate stack potential V_{di} .

2.10.2.1 Integration Constant $\varphi_{s,0}$

There are traditional ways of extracting the integration constant in (2.76), namely the zero-bias surface potential $\varphi_{s,0}$ [3]. We outline here two new methodologies for obtaining this integration constant.

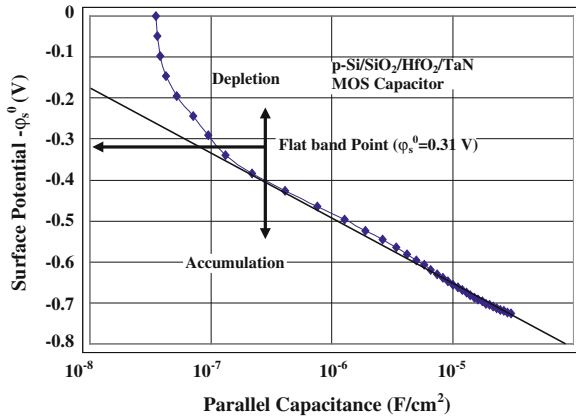
New Integration Constant Extraction Method 1.—The 100 kHz capacitance–voltage curve of Fig. 2.33 was integrated, according to (2.75), to obtain the surface potential. First the parallel capacitance, $C_p (=C_{sc} + C_{it})$, was calculated, using the relation:

$$\frac{1}{C_p} = \frac{1}{C} - \frac{1}{C_{di}} \quad (2.77)$$

Subsequently, C_p is plotted as a function of $\varphi_s - \varphi_{s,0}$, as illustrated by Fig. 2.40. It can be observed in Fig. 2.40, that the parallel capacitance C_p is very much an exponential function of the surface potential in the accumulation regime, and can be expressed as:

$$C_{p,acc} = \alpha_{acc} \exp(\beta_{acc} \varphi_s) \quad (2.78)$$

Fig. 2.40 Plot of the experimental parallel capacitance $C_p (=C_{sc} + C_{it})$ versus the surface potential $\varphi_s - \varphi_{s,0}$ (where $\varphi_{s,0}$ is the surface potential at zero bias). This plot is obtained using the 10 kHz C–V of Fig. 2.33 [72]



In (2.78), the pre-exponential constant α_{acc} is the value of $C_{p,acc}$ for $\varphi_s = 0$, and β_{acc} is the exponential constant of the surface potential φ_s .

In the next step, the flat-band space-charge capacitance, $C_{sc,fb}$, is calculated, using the bulk value of the acceptor density in the following relation, cf. (12):

$$C_{sc,fb} = \sqrt{q\epsilon_s N_A \beta} \quad (2.79)$$

Under the flat-band condition, $\varphi_s = 0$, the mathematical relation (2.12) simplifies exactly to (2.79). In Fig. 2.40, where $C_p = C_{sc,fb}$, there, the corresponding value of $\varphi_s - \varphi_{s,0}$ is $-\varphi_{s,0}$. As illustrated in Fig. 2.40, the value of the zero-bias surface potential is 0.31 V.

New Integration Constant Extraction Method 2.—The above methodology 1 is susceptible to errors if the surface doping density deviates significantly from the initial substrate value, and/or there is significant contribution from the interface traps to the flat-band parallel capacitance at the measurement frequency. A second option for calculating $\varphi_{s,0}$ would be to make use of the following empirical relation:

$$\frac{C_p}{\alpha_{acc}} = \sqrt{2} \quad (2.80)$$

In Fig. 2.40, where the above empirical relation holds, the corresponding value of $\varphi_s - \varphi_s^0$ is $-\varphi_s^0$. The genesis of (2.80) could be outlined in the following manner: If one compares (2.78) to (2.17), then, it would emerge that:

$$\alpha_{acc} = \sqrt{\frac{q\epsilon_s N_A \beta}{2}} \quad (2.81)$$

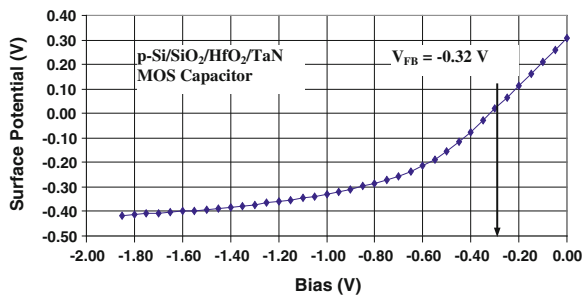
Comparison of (2.81) with (2.79) would yield the relation (2.80), if one neglects the contribution of the interface traps to the parallel capacitance density at flat-band:

$$C_{p,fb} = C_{sc,fb} + C_{it,fb} \cong C_{sc,fb}$$

Since the relation (2.80) is a ratio, any variation in the doping density and/or any contribution by the interface traps to the flat-band capacitance would affect both C_p and α_{acc} in the same manner; hence, methodology 2 is more immune to a variation in the value of the doping density in the semiconductor sub-surface or to a contribution to $C_{p,fb}$ from D_{it} .

Figure 2.41 presents the experimental surface potential versus the bias relation. The flat-band voltage, V_{FB} , is the value of the bias, corresponding to $\varphi_s = 0$, which came out to be -0.32 V, cf. Fig. 2.40. In Fig. 2.40, both the options for the extraction of $\varphi_{s,0}$ yielded nearly the same result, perhaps because there was no additional doping of the silicon sub-surface, and the interface trap density near the flat-band point was below $10^{11} \text{ cm}^{-2} \text{ V}^{-1}$, cf. Sect. 2.10.3.2. The surface potential plot $\varphi_s(V)$ of Fig. 2.41 can be considered to be reliable and accurate inside the accumulation regime, which covers much of the plot in Fig. 2.33. A small part of

Fig. 2.41 The Experimental surface potential versus the bias plot for the MOS capacitor of Figs. 2.40 and 2.33 [72]



the plot, i.e. in the voltage range of -0.32 to 0 V, in Fig. 2.41 is in the depletion regime. The accuracy of the surface potential in the depletion regime will depend upon the magnitude of the interface trap density in this range; the interface trap density extracted from the conductance technique was observed to be below $10^{11} \text{ cm}^{-2} \text{ V}^{-1}$ in this range, cf. Sect. 2.10.3.2.

The gate stack capacitance C_{di} used to be considered a constant and invariant of the bias and the signal frequency. This assumption is valid if the gate stack is a perfect dielectric and is devoid of any traps and therefore is devoid of any trap capacitance. This was the case for the single SiO_2 gate dielectric. However, as has been pointed out in detail, the high- k gate stack capacitance may contain a significant contribution from the gate stack traps in both deep accumulation and in deep inversion; in that case C_{di} would be a function of the bias V and the frequency f . The relation (2.75) would still be valid, as there is no assumption made in (2.76) that C_{di} is not a function of the bias V . However, there would be serious implementation problems. Since, the gate stack capacitance density C_{di} would keep increasing with the intensity of accumulation because of contribution from additional gate stack traps, it is not clear how C_{di} could be determined. Secondly, it is not clear how calculations would be made using the relation of (2.75) with a gate-voltage variant C_{di} .

2.10.3 Different Techniques for Trap Parameter Extraction

As mentioned in Chap. 1, the classical MOS trap parameter extraction techniques were developed for and applied to (i.e. were suitable for) non-leaky SiO_2 gate dielectrics. The mainstream techniques, responsible for the remarkable success of the SiO_2 gate technology, were the Low-High Frequency Capacitance Technique and the Conductance Technique [3, 4], both of which required the ability to obtain the quasi-static or the static C - V characteristic and reliable measurement of the G - V characteristic over a wide frequency range. Other (significantly less reliable and/or versatile) techniques included the Terman Technique and the Charge Pumping Technique. As mentioned already, leaky high- k gate stacks enormously complicate trap parameter extraction because the quasi-static, leave alone the static, C - V characteristic cannot be obtained under the normal conditions and because the

high- k gate stacks host a very complicated and still-unknown network of traps. As is outlined below, the low–high frequency technique and the conductance technique can still be used in a significantly truncated form and the information these techniques yield is also greatly reduced; moreover, the classical procedure for data analysis has to be modified (as detailed in [Sects. 2.10.3.1](#) and [2.10.3.2](#)) to make both these techniques usable in the case of the high- k gate stacks.

Inversion Capacitance—The main problems with the high gate stack leakage current are two-fold, namely, (a) the gate leakage drains the inversion layer, preventing its formation at the semiconductor surface and the measurement of the inversion capacitance, whereas (b) the high direct conductance submerges the alternating conductance measured at low and even moderate frequencies, thereby significantly restricting the scope of the conductance technique. There exist in principle three options for obtaining the inversion capacitance–voltage characteristic when this cannot be obtained under the normal conditions. Each of these three options enhances the minority carrier generation rate by orders of magnitude over its thermal value in the dark. An inversion layer will form at the semiconductor surface if the minority carrier supply rate exceeds its drainage from the surface by the gate stack leakage [51]. The rate of supply of the minority carriers to the surface will be proportional to its generation rate.

1. Option 1 involves measuring the admittance-voltage-frequency characteristics under illumination; the photo-generation enhances the electron–hole pair generation rate. This option is described in [Sect. 2.10.3.3](#).
2. Option 2 involves the measurement of the admittance-voltage-frequency characteristics at elevated temperatures; this methodology may be referred to as the High Temperature Admittance Technique. The genesis of this option is the enhanced minority carrier generation rate at higher temperatures. The minority carrier generation rate is proportional to the intrinsic carrier density $n_i = (N_c N_v)^{1/2} \exp(-E_G/2kT)$; therefore, higher temperatures translate into higher n_i and therefore higher minority carrier generation rate (N_c , N_v are the effective density of states in the conductance, valence band, respectively). The High Temperature Admittance Technique has been rarely used in practice. Consequently, it has not developed into a mature trap extraction technique. Moreover, it is not clear how much the enhanced minority carrier generation rate will mitigate the inversion layer drainage, as higher temperature can translate into a higher gate leakage rate.
3. Option 3 involves measurement of the MOS admittance-voltage-frequency characteristics using the MOSFET configuration and formation of the inversion layer by minority carrier injection from the source, to compensate for the gate stack leakage. Much like the option 2, this technique remains to be developed into a mature one with all the issues carefully analyzed and a comprehensive methodology is available for data analysis.

The problem of the high direct conductance submerging the alternating conductance can perhaps be solved if the small signal conductance could be measured

at frequencies higher than what has been used so far, namely frequencies higher than 1 MHz—say in the range of 1 MHz–10 GHz—to compensate for the loss of the range of, say, 1 Hz–10 kHz (The small signal conductance increases with the signal frequency). Attempts have been made to measure the conductance in this range (HF, UHF, Microwave), but serious measurement problems including those of the appropriate sample configuration, lead impedance, etc. are still to be overcome.

Terman Technique—Once the Low–High Frequency technique matured, the Terman Technique [61] (also referred to as the High Frequency Capacitance Technique) was seldom used in the case of the SiO₂ gate dielectrics on account of its unreliability and inaccuracy. In brief, it consists of the following steps: (1) The high frequency C–V characteristic is measured. (2) The ideal high frequency C–V characteristic is calculated. (3) The voltage shift between the measured and the calculated C–V curves, δV , is calculated as a function of the gate bias V_G and subsequently transformed into a function of the surface potential ϕ_s , using the calculated $\phi_s(V)$ relation. (4) The interface trap density is calculated from the differential $d(\delta V)/d\phi_s$, using the relation $D_{it} = (C_{di}/q) d(\delta V)/d\phi_s$. (5) The trap energy is calculated from the calculated value of ϕ_s . The Terman Technique has been applied to high-k gate stacks in some investigations, but it is highly unreliable because of the following reasons: (a) The potential δV , which is a part of the potential across the high-k gate stack, is small, when EOT is small, and is therefore vulnerable to error. (b) A significant inaccuracy is possible in the calculated high frequency C–V characteristic because of uncertainty in the value of the doping density. (c) As explained in Sect. 2.9.1.1, the depletion regime where the Terman Technique applies is tiny, in the case of the ultrathin gate stacks, with a flat profile which will promote errors in the value of δV .

Charge Pumping Technique—All the techniques discussed above are steady state small signal techniques. In contrast, the Charge Pumping Technique [70, 71] is a non-steady-state technique, in which while applying a reverse bias to the source/substrate and drain/substrate diodes, the gate is periodically pulsed between accumulation and strong inversion. The substrate direct current, which is called the charge pumping current, and is monitored over the duration of the pulse, originates from the exchange (emission and capture) of electrons and holes between the traps and the conduction band and the valence band of the semiconductor. All the non-steady-state and the transient techniques involve measurements during the period of trap relaxation in response to a change in the trap occupancy caused by a scanning of the imref through the trap levels; a transient capacitance technique measures the transient capacitance, whereas the Charge Pumping Technique measures the transient charging current flowing to adjust the trap charge due to a change in its occupancy.

The charge pumping model [71] indicates the current to depend upon a host of parameters including the trap density and the trap capture/emission cross-section. The charge pumping current I_{CP} has been observed to be proportional to the pulse frequency and the gate area, but its mathematical relation to many other parameters of the model has not been established. The usual variable measurement

parameters are the pulse frequency, the pulse width, and the pulse rise/fall times, the most potent among which is the pulse frequency. As the trap response time has to match the pulse frequency, in principle, it may be possible to access traps at different locations inside the gate stack and traps with widely varying capture/emission cross sections. However, in practice, there are several issues connected to the application of this technique to the high-k gate stacks:

1. In reality, the Charge Pumping Technique involves a complicated process, particularly in the case of the high-k gate stacks. The connection of the charge pumping current to the trap density (D_{it}), the trap capture cross-section (σ_t), and the trap location (x_t) in the gate stack is not clearly established yet; there are several ambiguities in the charge pumping models developed so far.
2. Any of the three trap parameters— D_{it} , σ_t , or x_t —cannot be determined reliably, unless the other two parameters are known from independent measurements. Since the latter is seldom the case, the values of the other two trap parameters are assumed. For example, often, the trap capture/emission cross-section (σ_t) is assumed to be of the order of 10^{-15} cm^{-2} ; but, experiments indicate that the capture cross-section of traps in the high-k gate stacks can vary by orders of magnitude [17], also see Chap. 8.
3. It is not clear what exactly the parameter D_{it} in the model for the charge pumping current I_{CP} represents.
4. The trap energy E_t cannot be extracted from this technique.
5. It is unclear what the effect is of the large gate leakage current on the reliability of this technique.
6. The charge pumping current has a clearer interpretation when the trap density is not a function of energy, the trap capture/emission cross-sections for electrons/holes are same for all the traps which respond to the pulse frequency, and all the traps which emit or capture electrons/holes are located on the same plane in the gate stack. In a high-k gate stack, the traps are located at various planes throughout the gate stack with capture/emission cross-sections varying over several orders of magnitude. So, it is possible that traps at different locations, with varying cross-sections and varying densities respond to a particular pulse frequency. What is the expression for the charge pumping current in such a case; can any trap parameter be extracted from this current?

2.10.3.1 Low-High Frequency Capacitance Technique

The high frequency C–V and the low frequency C–V are used in this technique. The low–high frequency capacitance technique consists of the following steps [3].

1. For any value of the bias in the depletion regime, the space charge capacitance density C_{sc} is calculated from the high frequency capacitance density C_{hf} , cf. (2.7) and Fig. 2.4a, whereas the parallel capacitance density $C_p = C_{sc} + C_{it}$ is

calculated from the low frequency capacitance density C_{lf} , cf. (2.6) and Fig. 2.4b, using the extracted value of the gate stack capacitance density C_{di} , cf. Sect. 2.10.1.1.

2. The experimental C_{sc} is subtracted from the experimental C_p to obtain the interface trap capacitance density C_{it} , from which the interface trap density D_{it} is extracted according to (2.6).
3. The surface potential is obtained from the bias value as outlined in Sect. 2.10.2, from which the trap energy E_{it} is extracted according to the following relation, for p-type semiconductor:

$$E_{it} = E_v + q(\phi_p + \phi_s) \quad (2.82)$$

4. The low–high frequency capacitance technique yields only the interface trap density distribution $D_{it}(E_{it})$; no trap time constant or capture cross-section data are obtained.
5. To obtain the $D_{it}(E_{it})$ distribution in the weak inversion regime, one has to use a space charge capacitance density calculated using a value for the doping density; this procedure is vulnerable to the uncertainty in the value of the doping density.

The low–high frequency capacitance technique is difficult to implement in the case of the leaky ultrathin gate stacks because of the following reasons. Generally the lowest frequency at which the C–V can be reliably measured is 1 kHz or so, which is far above the equilibrium frequency, and the highest frequency at which the C–V can be reliably measured is 100 kHz or so, which is below what can qualify as the high frequency. Consequently, even in the depletion regime, the low–high frequency capacitance technique can yield only a fraction of the total trap density at and around the Si/SiO₂ interface. However, it may be still be used only for an indicative comparison between samples.

2.10.3.2 Conductance Technique

As in the case of the leaky ultrathin gate stacks, the quasi-static C–V cannot be measured, the only reliable techniques, that are available for obtaining the trap parameters (trap density D_t , trap energy E_t , trap capture cross-section σ_t , and trap location x_t inside the gate stack), are the conductance technique [3, 4] and the charge-pumping technique [70, 71]. The conductance technique remains the most reliable technique for yielding D_t , E_t , and σ_t of traps in the majority carrier band-gap half, i.e. under the depletion condition [3]. To use the conductance technique, one needs to obtain the surface potential ϕ_s fairly accurately and the parallel conductance G_p , see Figs. 2.29 and 1.1. In principle, the measured total conductance G_m needs to be corrected for the series resistance R_s and the direct conductance G_{dc} . The direct conductance G_{dc} can be obtained by differentiating the measured I–V (direct current–voltage) characteristic [51]. However, no

satisfactory technique exists for obtaining the series resistance R_s in the case of a leaky high-k gate stack. The main reason is that the passive regions of the device cannot be represented by a lumped resistance, but, needs to be represented by impedances at several locations. In the case of non-leaky MOS devices with a single SiO_2 gate dielectric, the series resistance could be obtained from the real part of the impedance measured around 1 MHz in strong accumulation [3, 51].

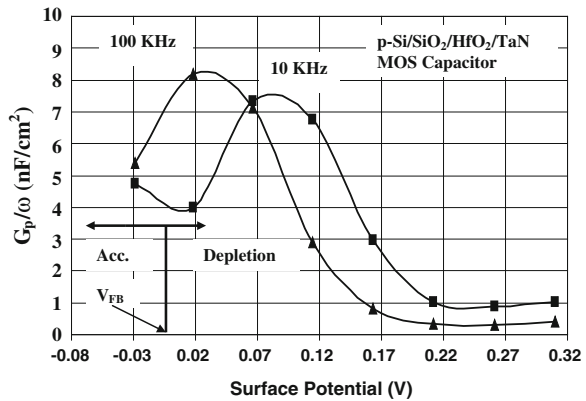
After the measured total parallel conductance G_m has been corrected for the series resistance R_s and the direct conductance G_{dc} , see Fig. 2.29, to obtain the total parallel conductance G_{ac} , the parallel conductance G_p is obtained using the following relation:

$$\frac{G_p}{\omega} = \frac{\omega(C_{di})^2 G_{ac}}{[(G_{ac})^2 + \omega^2(C_{di} - C)^2]} \quad (2.83)$$

Experimental G_p/ω is then plotted as a function of the surface potential ϕ_s , as depicted in Fig. 2.42, for 10 and 100 kHz, respectively, for the sample of Figs. 2.41 and 2.34. It may be noted that the curve, closer to the accumulation regime, represents the higher frequency (i.e. 100 kHz). As illustrated in Fig. 2.42, G_p/ω undergoes a peak, as ϕ_s is changed; the peak value of G_p/ω , $(G_p/\omega)_p$, is a measure of the trap density; while the corresponding value of ϕ_s , $(\phi_s)_p$, represents the trap's capture probability [3].

As already mentioned, when SiO_2 (grown by dry thermal oxidation) is the gate dielectric, it is safe to assume that traps exist mainly at or in the vicinity of the Si-SiO₂ interface. But, in the case of the high-k gate stacks, traps are likely to exist throughout the gate stack. Moreover, the trap density is perhaps lowest at the Si-SiO₂ (or Si-SiON) interface, and is higher or much higher elsewhere in the high-k gate stack [16]. So it is not necessary at all that only traps at the Si-SiO₂ interface will contribute to the observed conductance peaks in Fig. 2.42; the G_p/ω peaks in Fig. 2.42 will in general represent traps at some location x_t in the high-k gate stack. The trap density D_{it} can be calculated using the relation [3]:

Fig. 2.42 The experimental G_p/ω versus the surface potential under the depletion condition for the MOS capacitor of Figs. 2.31 and 2.32, for 10 and 100 kHz, respectively [72]



$$D_{it} = \frac{1}{f_D q} \left(\frac{G_p}{\omega} \right)_p \quad (2.84)$$

The parameter f_D represents the effect of the statistical fluctuation of the surface potential on G_p/ω [3]. The trap energy E_t can be calculated (for p-type silicon) according to (2.82). The trap's hole-capture cross-section can be extracted using the relation [3], cf. (2.62):

$$\sigma_t^h(x_t) p(x_t) = \frac{\omega}{f_\sigma v} \quad (2.85)$$

$\sigma_t^h(x_t)$ is the hole-capture cross-section of the trap at location x_t from the silicon surface, $p(x_t)$ is the hole density at x_t , v is the average thermal velocity of holes, and f_σ represents the effect of the statistical fluctuation of the surface potential on the trap time constant. In the case of single level interface traps at the Si/SiO₂ interface, $f_D = 0.5$ and f_σ is 1.0; otherwise, for a trap eigenenergy continuum at the Si/SiO₂ interface and statistical fluctuation of φ_s , the value of f_D is <0.5 , and can be as low as 0.15, while the value of f_σ is >1.0 and can be as high as 2.6 [3]. There is no analysis of how, in the case of high-k gate stacks, the statistical fluctuation of traps throughout the gate stack will affect the parameters f_D and f_σ . These two parameters are complex functions of the standard deviation, σ_s , of the surface potential [3].

A practical approach is outlined in the following for calculating the statistical parameters f_D and f_σ , which differs from that outlined by Nicollian and Brews [3]. Normally, the ratio $(G_p/\omega)/(G_p/\omega)_p$ is calculated from the plots of G_p/ω versus the small-signal frequency f , at 5 or 0.2 times the peak frequency f_p , at which the peak of G_p/ω occurs. Since in the case of the ultrathin high-k gate dielectric stacks, the frequency range for the conductance measurements is severely limited by the high gate dielectric leakage, one does not have enough frequency variation to obtain the peak and/or enough of the profile in the G_p/ω versus f plot. A solution for this problem is to use the value of G_p/ω in Fig. 2.42 at a surface potential, either $\Delta\varphi_s$ higher or lower than $(\varphi_s)_p$; $\Delta\varphi_s$ can be calculated using the relation:

$$\Delta\varphi_s = \frac{\ln 5}{\beta} \quad (2.86)$$

Once the ratio of $(G_p/\omega)/(G_p/\omega)_p$, is obtained at $(\varphi_s)_p \pm \Delta\varphi_s$, the standard deviation σ_s can be obtained from a numerical plot of $(G_p/\omega)/(G_p/\omega)_p$ versus σ_s , contained in [3]. Subsequently, f_D and f_σ can be obtained from numerical plots of f_D versus σ_s and f_σ versus σ_s , respectively, also provided in [3]. Tables 2.2 and 2.3 present the experimental values of σ_s , f_D , f_σ , D_t , E_t , and the trap's hole-capture probability, $\sigma_t p(x_t)$.

As (2.85) suggests, it is not possible to extract either the trap's hole-capture cross-section or the trap's location x_t in the gate stack, unless the other parameter is determined independently, e.g. from the charge-pumping technique. If the trap is located at the silicon surface ($x_t = 0$), i.e. it is an Si/SiO₂ interface trap, then the

Table 2.2 Experimental trap parameters

f (kHz)	$(G_p/\omega)/(G_p/\omega)_p$	σ_s ($1/\beta$)	f_D	f_σ	D_t ($\text{cm}^{-2} \text{V}^{-1}$)
10	0.803	1.86	0.250	2.400	1.75×10^{11}
100	0.829	2.05	0.238	2.425	2.00×10^{11}

Table 2.3 Experimental trap parameters

f (kHz)	$E_t - E_v$ (eV)	$\sigma_t^h p(x_t)$ (cm^{-1})	$\sigma_{it}^h(x_t = 0)$ (cm^2)	$\sigma_{it}^h(x_t = 0.4 \text{ nm})$ (cm^2)
10	0.31	2.62×10^{-3}	4.21×10^{-17}	1.05×10^{-15}
100	0.26	2.59×10^{-2}	4.95×10^{-17}	1.24×10^{-15}

values of the trap's hole-capture cross-section come out to be very small (of the order of 10^{-17} cm^2), as indicated in Table 2.2. In principle, a range as large as 10^{-12} – 10^{-18} cm^2 is possible for the trap capture cross-section [44], depending upon whether the trap is charge-wise neutral, or Coulomb-attractive (very large σ_t) or Coulomb-repulsive (very small σ_t). If on the other hand, the trap is located inside the gate stack, its hole-capture cross-section would be larger. For example, if we assume an x_t of 0.4 nm, and $(2\kappa_h)^{-1} = 0.16 \text{ nm}$, then the hole capture cross-sections are of the order of 10^{-15} cm^2 (typical values), as Table 2.3 indicates.

As Table 2.2 amply illustrates, if the effects of the statistical fluctuation of the surface potential are ignored, large errors will result in the values of the trap density and the trap's capture cross-section. In the literature, almost always, not only has this phenomenon been ignored in extracting the trap parameters from the conductance data, but single level traps have been assumed without any validation. Also, in most cases, only the trap density has been estimated from the conductance data, but not the trap energy or the trap capture cross-section. The main reason behind this deficiency has perhaps been the inability to extract the surface potential, in the absence of the quasi-static C–V data.

This brings us to the issue of the error in the surface potential data of Fig. 2.41 and the methodology outlined in Sect. 2.10.2 for extracting the surface potential in the absence of the quasi-static C–V characteristic. The extracted values of the surface potential, as explained earlier, are accurate in the accumulation regime. The magnitude of error in the surface potential in the depletion regime would increase as the surface potential moves away from the flat-band point and towards the weak inversion condition; the error would also depend on the magnitude of the trap density. The values of the trap density, as reflected in Table 2.2, are relatively low for a high-k gate stack, particularly, when one considers that these traps are close to the valence band edge, an energy range, where the trap density is likely to be significantly higher than the mid-gap trap density. One may bear in mind that, these values, of the trap density in Table 2.2, may not represent the total trap density, which decides the magnitude of the total potential across the gate stack. It is useful to note that, in a situation, where traps are present throughout the gate stack, the trap capacitance at the equilibrium frequency will reflect the total trap density, but the conductance at any frequency will reflect the trap density only at a certain x_t [72].

The experimental data of Figs. 2.33, 2.34, 2.40, 2.41 and 2.42 belong to the same sample; this sample belonged to a group of samples on wafer D-06 which had a graded SiO_2 layer (1–6 nm) grown in dry O_2 at 900 °C. Figures 2.43 and 2.44 illustrate the variation in the experimental interface trap density and the hole capture cross-section, respectively, with EOT, for different sets of such wafers (All these wafers had graded SiO_2 layer). Wafer D-04 had no HfO_2 layer; wafer D-06 and D-12 had 2 nm and wafer D-10 and D-14 had 3 nm thick HfO_2 layer. Wafers D-06 and D-10 had a post-deposition annealing (PDA) in O_2 at 500 °C for 1 min. The wafers of Fig. 2.31 are the same as those of Figs. 2.43 and 2.44. The 100 kHz conductance yielded the parameters of traps located in the range of 0.24–0.27 eV above the valence band edge E_v , while the 10 kHz conductance yielded the parameters of traps located in the range of 0.30–0.32 eV above the valence band edge E_v (see Table 2.3). The trap energy E_t and the hole capture cross-section σ_h did not vary significantly with the SiO_2 layer thickness or the HfO_2 layer thickness, suggesting that in the case of all the MOS capacitors (high-k gate stacks), we are

Fig. 2.43 Interface trap density versus EOT for p-Si/ SiO_2 / HfO_2 /TaN MOS capacitors on different graded- SiO_2 wafers with different sets of PDA and HfO_2 thickness [17]

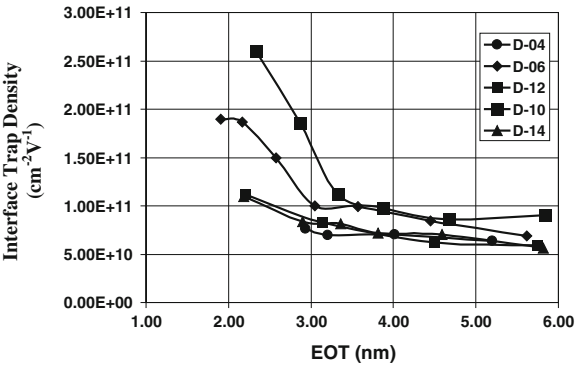
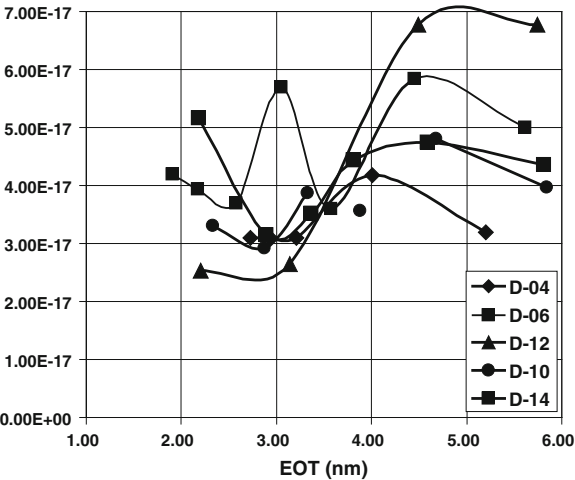


Fig. 2.44 Hole capture cross-section versus EOT for p-Si/ SiO_2 / HfO_2 /TaN MOS capacitors on different graded- SiO_2 wafers with different sets of PDA and HfO_2 thickness [17]



dealing with the same nature of traps. Figures 2.43 and 2.44 represent the interface state density at $E_v + 0.26 - 0.28$ eV.

The experimental data on the trap density, as illustrated by Fig. 2.43, may be summarized in the following:

1. The trap density increases for all wafers with decreasing SiO₂ thickness, except for wafer D-04 (SiO₂ gate dielectric).
2. The increase in the trap density for the thinner SiO₂ layer is significantly higher for wafers subjected to Post Deposition Annealing (PDA in oxygen at 500 °C for 1 min.).
3. The increase in the trap density is higher for the thicker HfO₂ layer, in the case of wafers undergoing PDA.

Following interpretations are possible from the results on the trap parameters in Figs. 2.43 and 2.44 and on the flat-band voltage profile as a function of EOT in Fig. 2.31:

1. That the trap density does not change significantly with decreasing SiO₂ thickness in the case of wafer D-04 (no HfO₂ layer) suggests that the increase in D_{it} in the case of wafers with the HfO₂ layer has something to do with the HfO₂ layer itself; also it has nothing to do with the TaN gate electrode.
2. This conclusion is reinforced by the fact that the flat-band interface charge density changes from a net negative value for the wafer D-04 (no HfO₂ layer) to a net positive value for the other wafers all of which have HfO₂ layer, cf. Fig. 2.31 and Sect. 2.8.1. The change in the sign of the flat-band interface charge with the induction of the HfO₂ layer suggests introduction of new defects into the intermediate SiO₂ layer and the Si-SiO₂ interface by the HfO₂ layer.
3. That the trap density in Fig. 2.43 and its increase for the thinner SiO₂ are higher for the thicker HfO₂ layer is another fact indicating the crucial role of the hafnia layer. Chapter 7 presents experimental data showing higher trap density in the intermediate SiO₂ layer as the HfO₂ layer thickness is increased.
4. The HfO₂ layer and its two interfaces with a huge flat-band charge density of $q \times 1.5 \times 10^{13} \text{ cm}^{-2}$ (see Fig. 2.31 and Sect. 2.8.1) represent an inexhaustible store of defects and source of contamination to diffuse into the neighbor SiO₂ layer and the Si-SiO₂ interface with a two orders of magnitude lower trap density. The defect concentration gradient across the HfO₂-SiO₂ layers is very steep; what exactly diffuses is not clear.
5. The increase in D_{it} for thinner SiO₂ can be explained by enhanced diffusion of impurity ions or atoms from the HfO₂ layer through the thinner SiO₂. It can also be explained by the influence of the SiO₂/HfO₂ interface on the generation of traps in the SiO₂ layer—due to mismatch in chemical bonding, in coordination, and in impurity or defect solid solubility; such a trap is likely to have a profile with a peak near the SiO₂/HfO₂ interface; so, D_{it} at the Si/SiO₂ interface will be higher for a thinner SiO₂ layer.

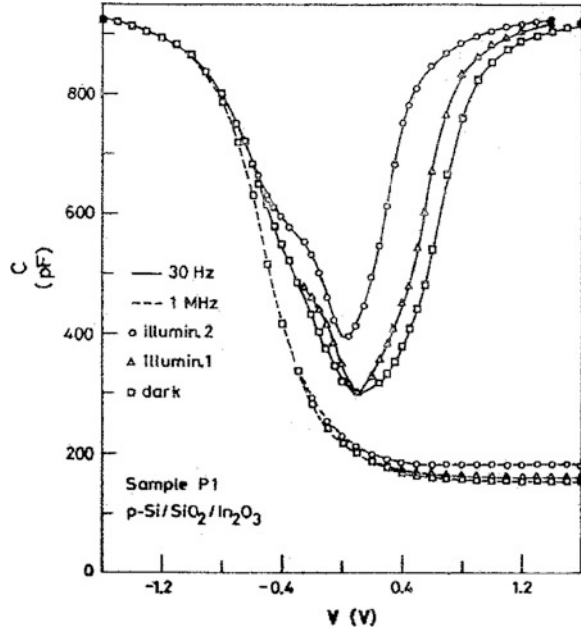
6. The formation of an interface dipole at the $\text{SiO}_2/\text{HfO}_2$ interface has been analyzed in depth in Chap. 6 and an interface dipole model has been presented based upon the difference in the areal density of oxygen between SiO_2 and HfO_2 , which results in the exchange of oxygen across the $\text{SiO}_2/\text{HfO}_2$ interface, culminating in an interface dipole. The traps of Fig. 2.43 may be the tail of this interface dipole.
7. Oxygen vacancies are generated in the hafnia layer; the released oxygen atoms may diffuse through the thinner silica layer into the silicon/silica interface [73] and induce new traps.
8. As oxygen diffuses through a thinner silica layer to the Si-SiO₂ interface; a new oxide layer grows at lower temperature (500–600 °C during HfO_2 deposition and/or PDA), resulting in a higher trap density. The source of oxygen could be the HfO_2 layer (oxygen released by generation of oxygen vacancies), the gas phase during the HfO_2 deposition, and/or the gas phase during the PDA.
9. The degradation of the HfO_2 gate stack has been analyzed in detail in Chap. 8 with the conclusion that the main seat of the stress-induced traps, which are the main agents of degradation, is inside the intermediate SiO_2 layer. The experimental data presented in Chap. 8 suggest that these traps/defects are positively-charged oxygen vacancies induced in the SiO_2 layer by the hafnia layer.

2.10.3.3 Photo-Admittance Technique

The photo-admittance technique was proposed by Poon and Card [74]; subsequently it was developed by Kar and Varma [75] with a complete methodology for parameter extraction. This technique involves measuring the MOS admittance (capacitance and conductance) as a function of bias at different frequencies and at different illumination levels, see the C–V characteristics in Fig. 2.45. In principle, this technique enables measurement of the equilibrium C–V characteristic including the inversion capacitance in the case of the leaky gate stacks and/or channel (semiconductor substrate) materials with a low minority carrier generation rate. Under illumination, the minority carrier generation rate is enhanced by many orders of magnitude over its thermal generation rate in the dark; this increased minority carrier generation rate enables build-up of the inversion layer in leaky gate stacks and/or in the case of channels with a low thermal generation rate (as in GaAs). This technique was applied by Kar and Varma [75] to non-leaky SiO_2 gate dielectrics on Si to extract the trap density and the capture cross-section in both halves of the band-gap, i.e. under both the depletion and the weak inversion conditions. An important point to note is that this is the only technique which allows determination of both the electron and the hole capture cross-sections and the trap density in both halves of the band-gap from the measured MOS conductance irrespective of whether the gate stack is leaky or not.

As has been discussed in Chap. 12 and also in the other chapters, high mobility channels have to be employed to enhance the drain current. Many of these high

Fig. 2.45 Capacitance–voltage characteristics of a p-Si/SiO₂/In₂O₃ MOS structure (sample P1) measured at 30 Hz and 1 MHz in the dark and under two different levels of illumination [81]

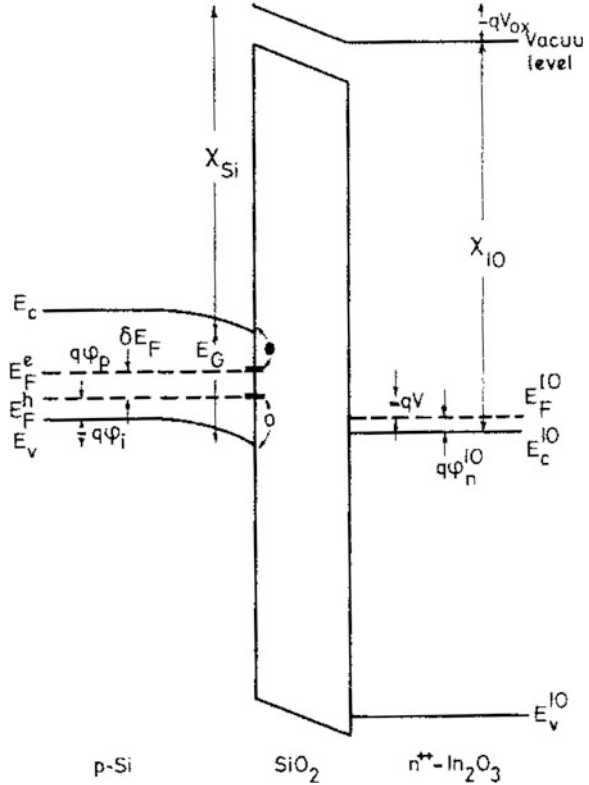


mobility channels have low minority carrier generation rate in the dark and consequently suffer from the inability to measure the inversion capacitance and the C–V characteristic in the weak and the strong inversion regimes. In such situations, the photo-admittance technique may be an effective aid. As will be explained later, the photo-capacitance or the photo-conductance can alone yield the trap density in both halves of the band-gap, but to extract the trap energy reliably, both the photo-capacitance and the photo-conductance have to be measured and analyzed [75].

Under illumination, the law of mass action will no longer hold and the magnitude of the imref (i.e. the quasi-Fermi level) separation will increase with the illumination level. Under the low-level injection, the majority carrier imref will remain the same as in the dark, while the minority carrier imref will move towards the minority carrier band edge, see Fig. 2.46. Consequently, under low-level injection, while the majority carrier density will remain the same as in the dark, the minority carrier density will increase by many orders of magnitude. Hence, the space charge capacitance under illumination $C_{sc,photo}$ will change from its value in the dark C_{sc} , and the mathematical relation of (2.12) has to be modified as indicated below, in which δE_{FS} is the imref separation [75]:

$$C_{sc,photo} = \left(\frac{q\epsilon_s N_A \beta}{2} \right)^{1/2} \frac{\left| 1 - e^{-\beta\phi_s} + n_0/p_0 (e^{\beta\phi_s} - 1) \exp\left(\frac{\delta E_{FS}}{kT}\right) \right|}{\left[(e^{-\beta\phi_s} + \beta\phi_s - 1) + n_0/p_0 (e^{\beta\phi_s} - \beta\phi_s - 1) \exp\left(\frac{\delta E_{FS}}{kT}\right) \right]^{1/2}} \quad (2.87)$$

Fig. 2.46 Energy band diagram of a p-Si/SiO₂/In₂O₃ MOS structure at a certain applied bias V . E_F^h and E_F^e are hole and electron imref; ϕ_i is the surface (interface) potential and ϕ_p is the Fermi potential; E_F^{IO} is the indium oxide Fermi level and E_v^{IO} and E_c^{IO} are the indium oxide valence and conduction band edge; ϕ_n^{IO} is the Fermi potential in the degenerate indium oxide [75]



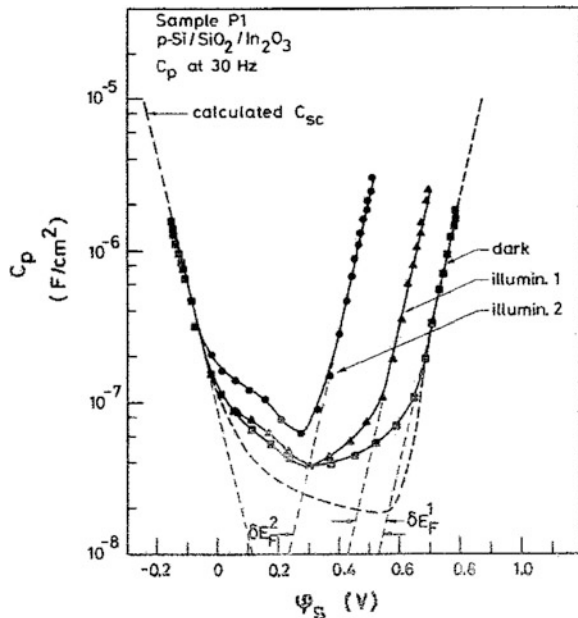
Extraction of the imref separation δE_{FS} is an important exercise in the photo-admittance technique. There are three options for determining the imref separation:

1. From the measured high frequency photo-capacitance in strong inversion. Under illumination, strong inversion will set in at a surface potential $\delta E_{FS}/q$ less than its value in the dark: $\phi_{s,inv,th,photo} = (\phi_{s,inv,th} - \delta E_{FS}/q)$. The surface potential at the onset of strong inversion under illumination $\phi_{s,inv,th,photo}$ can be determined from the high frequency photo-capacitance minimum in strong inversion; its deviation from its dark counterpart will yield the imref separation [75].
2. From the Berglund integral of the equilibrium photo-capacitance–voltage characteristic over strong accumulation to strong inversion, see (2.75), which will yield a surface potential $\Delta\phi_s$, the deviation of which from the band-gap equivalent voltage will yield the imref separation: $E_G - q\Delta\phi_s = \delta E_{FS}$ [75].
3. From the parallel shift of the parallel capacitance C_p (or the space charge capacitance C_{sc}) in strong inversion between its dark value $C_{p,inv}$ and that under illumination $C_{p,inv,photo}$. The experimental parallel capacitance C_p is extracted using (2.77) and the experimental surface potential ϕ_s is determined as outlined

in Sect. 2.10.2. The parallel capacitance C_p in dark as well as under illumination is plotted as a function of the surface potential φ_s , as is illustrated in Fig. 2.47. In strong inversion, the parallel capacitance, $C_p = C_{sc} + C_{it}$, generally reduces to the space charge capacitance C_{sc} . Figure 2.47 demonstrates the experimental C_p to be an exponential function of φ_s , in both dark and under illumination, as could be expected from (2.12), (2.17), and (2.87). As is indicated in Fig. 2.47, the parallel shift along the φ_s axis between the illuminated and the dark characteristic yields the imref separation δE_{FS} .

The trap parameters—the trap density D_{it} , trap energy E_t , and the capture cross-section of the trap σ_h/σ_e —are determined using the procedure outlined in Sects. 2.10.3.1 (Low–High Capacitance Technique) and 2.10.3.2 (Conductance Technique), with the following modifications. In the dark, interface trap recombination is dominated by exchange of carriers by traps at the Fermi level with the majority carrier band, even in the weak inversion regime [3]. However, under illumination as illustrated in Fig. 2.46, carrier exchange between interface traps and both the conduction and the valence bands are possible. Before the trap energy or the trap cross-section can be determined, one needs to establish whether the exchange of holes between the traps at the hole imref $E_{FS,h}$ and the valence band or the exchange of electrons between the traps at the electron imref $E_{FS,e}$ and the conduction band is dominating the trap recombination process, cf. Fig. 2.46. If $p_s\sigma_h$ is $\gg n_s\sigma_e$, then the interface trap recombination process is dominated by the hole capture by traps at $E_{FS,h}$; in this case, the trap energy is given by (2.82). On the other hand, if $p_s\sigma_h$ is $\ll n_s\sigma_e$, then the interface trap recombination process is

Fig. 2.47 Experimental C_p as a function of the surface potential of the same device P1 as in Fig. 2.45 in dark and two different levels of illumination. The broken line represents the calculated low frequency space charge capacitance density as a function of the surface potential in the dark condition [81]



dominated by the electron capture by traps at $E_{FS,e}$; in that case the trap energy is given by:

$$E_t = E_v + q(\phi_s + \phi_p) + \delta E_{FS} \quad (2.88)$$

To determine the capture cross-section from the conductance data by the procedure outlined in Sect. 2.10.3.2, one has to confirm whether a particular G_p/ω versus the ϕ_s characteristic represents carrier exchange with the valence band or with the conductance band. Unless the hole and the electron capture cross-sections are very unequal, generally, for p-type semiconductor, carrier exchange with the valence band will prevail in the depletion regime, and with the conduction band in the weak inversion regime, and vice versa for the n-type semiconductor. For a p-type semiconductor, if the hole exchange with the valence band dominates, then the position of the G_p/ω peak— $(\phi_s)_p$ —would move towards more positive values of ϕ_s with decreasing frequency, and towards more negative values if electron exchange with the conduction band dominates.

2.11 A Fundamental Basis for the Ultimate EOT

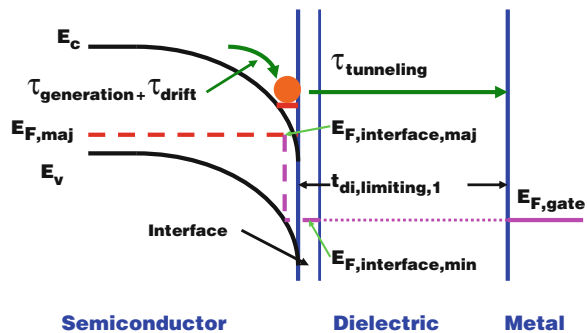
The issue of the lowest value of the EOT possible for a gate stack of an MOSFET has engaged our attention over 4 decades or longer. In olden times the question asked was how small the thickness of the SiO_2 gate dielectric could be in principle? The prevailing wisdom for a long time was that the gate dielectric had to be thick enough to prevent a direct tunneling current flowing through it. Mead, in spite of being a keen reader of the times to come, could only predict an ultimate thickness of 5 nm for the SiO_2 layer in 1972 [76]. Nicollian and Brews, notwithstanding the knowledge of the MOS basics they had, could not contemplate even in 1982 that tunneling SiO_2 layers would ever see the light of day as gate dielectrics in MOSFETs and considered the thin tunnel SiO_2 layers, leave alone the ultrathin ones, to be useless [3]. The current wisdom envisages the ultimate EOT to be around 0.5 nm; this prediction is based upon the gate leakage current and not so much on a more fundamental point of physics. It is interesting to note that the two most frequent parameters on which the ultimate EOT has been based are the gate leakage current and the sensitivity of lithography. Both these parameters have witnessed several generations; each generation has experienced a scaling down. Is there a more basic or fundamental feature which has an important bearing on the lowest EOT possible for the gate stack? In the following we will attempt to suggest that the EOT threshold for the conversion of the tunnel MOS structure to a Schottky barrier is such a feature.

Kar and Dahlke [51] and Kar [77] had classified the MOS Tunnel Diodes into two groups: (1) *Intermediate Tunnel MOS Structure* in which the minority carrier imref at the semiconductor surface was pinned to the metal Fermi level, but the majority carrier imref remained pinned to the majority carrier imref in the

semiconductor bulk; (2) *Schottky Tunnel MOS Structure* in which both the majority carrier as well as the minority carrier imref at the semiconductor surface are pinned to the metal Fermi level. In the intermediate tunnel MOS structures, the occupancy of the states at the Si-SiO₂ interface is determined by the majority carrier imref in the bulk semiconductor, whereas in the Schottky tunnel MOS structures, the occupancy of the states at the Si-SiO₂ interface is determined by the metal Fermi level. There exist two threshold values of the EOT at which the transformation from the thick (non-leaky) MOS structure to intermediate tunnel MOS structure and from the intermediate tunnel MOS structure to the Schottky tunnel MOS structure, respectively, occur [77].

For an EOT lower than the $EOT_{\text{threshold,min}}$, the time the minority carriers take to reach the semiconductor surface from the semiconductor bulk ($=\tau_{\text{generation}}+\tau_{\text{drift}}$) exceeds their time of tunneling from the semiconductor surface to the metal ($\tau_{\text{tunneling}}$), as a result of which the minority carriers are drained away to the metal, no inversion layer forms at the semiconductor surface, and the minority carrier imref at the semiconductor surface gets pinned to the metal Fermi level, as illustrated in Fig. 2.48 [78]. The minority carrier generation time depends on the semiconductor substrate and its quality; in the case of device quality silicon, a typical value of $\tau_{\text{generation}}$ could be 10^{-5} s. After generation in the semiconductor bulk (neutral region), which is the main source of the supply of the minority carriers to the semiconductor surface, the minority carriers traverse the space-charge region by drift to reach the semiconductor surface; an estimate of this drift time could be 10^{-12} s, as suggested in Fig. 2.48. Hence, as indicated in Fig. 2.48, the conversion of the thick MOS to the intermediate tunnel MOS will occur at the threshold EOT of $EOT_{\text{threshold,min}}$, for which the minority carrier tunneling time $\tau_{\text{tunneling}}$ becomes much smaller than their generation time. Naturally the value of the $EOT_{\text{threshold,min}}$ will depend upon the composition of the gate stack, the

Fig. 2.48 Energy band diagram across semiconductor/dielectric/metal structure illustrating minority carrier interface imref pinning and the corresponding EOT threshold: $EOT_{\text{threshold,min}}$



$$\tau_{\text{drift}} = 10^{-5}/10^7 \text{ s} = 10^{-12} \text{ s} \quad \tau_{\text{generation}} + \tau_{\text{drift}} \approx \tau_{\text{generation}} = 10^{-5} \text{ s} \quad \tau_{\text{tunneling}} \ll 10^{-5} \text{ s} \Rightarrow \text{No strong inversion layer forms in an MOS structure ; } E_{F,\text{interface,min}} \text{ pinned to } E_{F,\text{gate}}; \text{ Onset of direct tunneling.}$$

semiconductor substrate and its quality, etc.; in the case of the SiO_2 gate dielectric, it was observed to be around 3.3 nm for a device quality silicon [51, 77].

For an EOT lower than the $\text{EOT}_{\text{threshold,maj}}$, the time the majority carriers take to reach the semiconductor surface from the semiconductor bulk ($=\tau_{\text{relaxation}} + \tau_{\text{drift}}$) exceeds their time of tunneling from the semiconductor surface to the metal ($\tau_{\text{tunneling}}$), as a result of which the states on the semiconductor surface exchange carriers faster with the metal than with the semiconductor bulk, the occupancy of states at the semiconductor surface is determined by the metal Fermi level, and the majority carrier imref at the semiconductor surface gets pinned to the metal Fermi level, as illustrated in Fig. 2.49. When, this happens, the isolation of the gate electrode from the semiconductor surface and the channel (the insulated gate character) completely disappears and no surface inversion or accumulation are possible. Consequently, with such a gate stack (i.e. with a Schottky tunnel MOS), the device will cease to function as an MOSFET and will be transformed to a Schottky barrier gate FET. In other words, $\text{EOT}_{\text{threshold,maj}}$ is the lowest value the EOT of an MOSFET gate stack can have. In a Schottky barrier gate FET, no accumulation or inversion layer is possible; the semiconductor surface can only be under the depletion condition.

As indicated in Fig. 2.49, the source of supply for the majority carriers is the bulk semiconductor relaxation (given by $\tau_{\text{relaxation}}$); to reach the semiconductor surface, the majority carriers traverse the space-charge layer by drift (given by τ_{drift}). Estimates of the different components of the total time $-\tau_{\text{relaxation}} + \tau_{\text{drift}}$ —it takes the states inside the conduction/valence band at the semiconductor surface to exchange carriers with the semiconductor bulk are indicated in Fig. 2.49 for silicon MOSFETs. For a semiconductor surface state inside the band-gap, an additional time—the recombination time $\tau_{\text{R,maj}}$ —has to be added to $-\tau_{\text{relaxation}} + \tau_{\text{drift}}$ —to obtain the total exchange time. To estimate the threshold EOT— $\text{EOT}_{\text{threshold,maj}}$ —we need to consider the smallest time the state at the semiconductor surface needs to communicate with the semiconductor bulk.

Fig. 2.49 Energy band diagram across semiconductor/dielectric/metal structure illustrating majority carrier interface imref pinning and the corresponding EOT threshold: $\text{EOT}_{\text{threshold,maj}}$

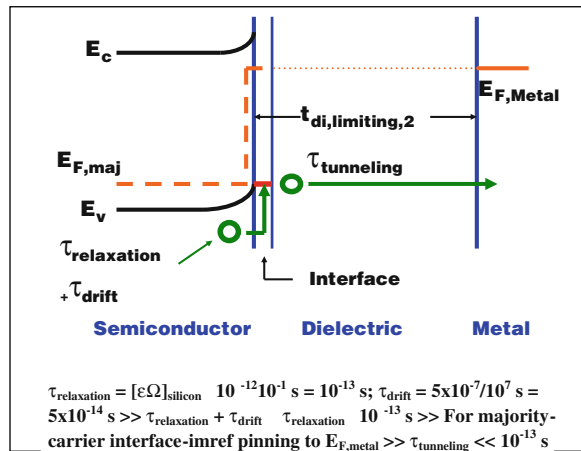
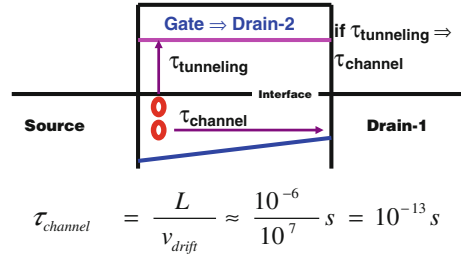


Fig. 2.50 A schematic of the MOSFET illustrating the competition between the gate metal and the regular drain to capture (by tunneling and by drift, respectively) the carrier injected by the source. The time to traverse the channel, τ_{channel} , is estimated assuming a channel length of 10 nm



The estimated times in Fig. 2.49 suggest that we basically need to compare the semiconductor relaxation time with the majority carrier tunneling time. The semiconductor relaxation time will depend upon the semiconductor permittivity and resistivity; the majority carrier tunneling time will depend upon the composition of the gate stack and the EOT. When the majority carrier tunneling time becomes smaller than the semiconductor relaxation time, then the intermediate tunnel MOS transforms into a Schottky tunnel MOS, the majority carrier surface imref gets pinned to the metal Fermi level, and the gate stack is no longer functional for the MOSFET operation. The threshold EOT for which this occurs, i.e. $\text{EOT}_{\text{threshold,maj}}$, is the lowest value of EOT possible for the gate stack for MOSFET operation, unless some other fundamental consideration yields a higher value for the minimum EOT. An experimental value for the $\text{EOT}_{\text{threshold,maj}}$ is not available at this time; in the case of the SiO_2 single gate dielectric, $\text{EOT}_{\text{threshold,maj}}$ should be less than 0.8 nm, as Intel reports MOSFET operation for an SiO_2 gate dielectric of 0.8 nm [79]. Muller reports that some 0.7 nm of SiO_2 is necessary for a meaningful tunnel barrier to take shape [80].

A basic consideration which may assign a higher value for the ultimate EOT than the $\text{EOT}_{\text{threshold,maj}}$ is illustrated in Fig. 2.50. If the time to traverse the channel exceeds the time of tunneling through the gate stack, then the carriers injected by the source would rather end up in the metal electrode than in the customary drain; in other words, the gate metal will become a secondary or even primary drain, and the MOSFET will cease to function. The time estimated for the channel traverse time, τ_{channel} , i.e. 10^{-13} s, in Fig. 2.50 is about the same as the semiconductor relaxation time estimated in Fig. 2.49. The channel traverse time would be determined by the channel length, the electric field along the y axis, and the channel mobility.

2.12 Summary

The material presented in this chapter can be classified into two groups—one set for the MOS/MOSFET devices with the SiO_2 single gate dielectric and the other set for the same with the high-k gate stacks. What comes out clearly and strongly is the large difference in the nature, behavior, and the characteristics of the two sets;

almost everything deviates strongly—be it the energy band profile, the circuit representation, the traps, the parameter extraction, and the mathematical relations for the channel parameters, for the gate stack charges, and for the gate stack potentials. In the case of the high- k gate stacks, the channel parameters—the drain current, the channel conductance, and the transconductance—are severely degraded by the non-ideal factors of the work-function anomaly, the high- k gate stack charges, and the non-saturating inversion surface potential; in addition, the drain current versus the drain voltage relation becomes more non-linear even in the triode regime. The energy band profile across the high- k gate stack is enormously complicated by its composition—two bulk dielectrics (intermediate layer—IL—and the high- k layer) and three transition layers with varying composition, varying band-gap, and varying permittivity respectively at the semiconductor/IL, IL/high- k , and high- k /metal interfaces. The corresponding circuit representations and the charge-potential relations are equally complicated; in fact, it is not possible to represent the transition region of varying permittivity with a finite number of circuit elements. The theoretical treatment and the mathematical relations are complicated also by the quantum-mechanical phenomena. The analysis presented in this chapter on the latter suggest a strong dilution of the carrier confinement in the strong inversion and the accumulation layers by wave function penetration into the gate stack, tunneling through the gate stack, and the metal/semiconductor wave function mixing. The concept of the pseudo-Fermi potential and the Fermi occupancy of traps inside the gate stack have been analyzed; this analysis suggests that it is likely that all the traps communicate and exchange carriers with either the semiconductor or the metal surface in accumulation and in strong inversion if the EOT is <1.0 nm and the ac signal frequency is <1 MHz. In such a case, for a measurement frequency of 1 MHz or less, the MOS capacitance in accumulation or in strong inversion will exceed the dielectric capacitance of the gate stack, since it will be augmented by the charging capacitance of the gate stack traps. An erroneous and false value of EOT will result if the same is extracted from this capacitance. Moreover, the augmentation of the dielectric capacitance by the charging capacitance of the gate stack traps can explain the huge frequency dispersion of the accumulation or the strong inversion capacitance observed particularly in the case of the high-mobility channels. The C – V characteristic of the ultrathin gate stack deviates strongly from the classical form. The classical MOS parameter extraction techniques do not apply in the case of the ultrathin gate stacks and need to be modified; one reason for this is the inability to apply the quasi-static C – V technique on account of the high gate leakage current.

References

1. A.S. Grove, *Physics and Technology of Semiconductor Devices* (Wiley, New York, 1967)
2. S.M. Sze, *Physics of Semiconductor Devices* (Wiley, New York, 1981)
3. E.H. Nicollian, J.R. Brews, *MOS Physics and Technology* (Wiley, New York, 1982)

4. E.H. Nicollian, A. Goetzberger, The Si-SiO₂ interface—electrical properties as determined by the MIS conductance technique. *Bell Syst. Tech. J.* **46**, 1055 (1967)
5. E.H. Rhoderic, R.H. Williams, *Metal-Semiconductor Contacts* (Clarendon Press, Oxford, 1988)
6. S.M. Sze (ed.), *Modern Semiconductor Devices* (Wiley, New York, 1998)
7. H. Lueth, *Surfaces and Interfaces of Solid Materials* (Springer, Berlin, 1995)
8. C.G.B. Garrett, W.H. Brattain, Physical theory of semiconductor surfaces. *Phys. Rev.* **99**, 376 (1955)
9. S. Kar, Interface charge characteristics of MOS structures with different metals on steam grown oxides. *Solid-St. Electron.* **18**, 723–732 (1975)
10. S. Kar, Determination of Si-metal work function differences by MOS capacitance technique. *Solid-St. Electron.* **18**, 169–181 (1975)
11. H.K.J. Ihantola, J.L. Moll, *Solid-State Electron.* **7**, 423 (1964)
12. S.R. Hofstein, F.P. Heiman, *Proc. IEEE* **51**, 1190 (1963)
13. C.G. Parker, G. Lucovsky, J.R. Hauser, Ultrathin oxide–nitride gate dielectric MOSFET's. *IEEE Electron Device Lett.* **19**(4), 106 (1998)
14. S. Kar, M. Houssa, S. Van Elshocht, D. Misra, K. Kita (eds.), Physics and technology of high-K materials IX. *ECS Trans.* **41**(3) (2011), ch. 8, ch. 7
15. S. Kar, S. Van Elshocht, D. Misra, K. Kita (eds.), Physics and technology of high-K materials X. *ECS Trans.* **41**(3) 2012
16. S. Kar, S. Rawat, *ECS Trans.* **16**(5), 443 (2008)
17. S. Kar, *ECS Trans.* **25**(8), 399 (2009)
18. T. Hori, *Gate Dielectrics and MOS ULSIs* (Springer, Berlin, 1997). (ch. 3)
19. S.M. Sze, *Modern Semiconductor Device Physics* (Wiley, New York, 1998). (ch. 3)
20. C.C. Hu, *Modern Semiconductor Devices for Integrated Circuits* (Prentice Hall, Upper Saddle River, 2009). (ch. 6)
21. S. Kar, *IEEE Trans. Electron Devices* **50**, 2112 (2003)
22. S. Kar, S. Rawat, S. Rakheja, D. Reddy, *IEEE Trans. Electron Devices* **52**, 1187 (2005)
23. R. Choi, S.J. Rhee, J.C. Lee, B.H. Lee, G. Bersuker, *IEEE Electron Device Lett.* **26**, 197 (2005)
24. G. Bersuker et al., *IEEE Trans. Device Mater. Reliab.* **7**, 138 (2007)
25. A. Toriuma, K. Kita, *ECS. Trans.* **19**(1), 243 (2009)
26. J.K. Schaeffer et al., *Appl. Phys. Lett.* **85**, 1826 (2004)
27. H. Park et al., *IEEE Electron Device Lett.* **26**, 725 (2005)
28. R. Xie, T.H. Phung, W. He, M. Yu, C. Zhu, *IEEE Trans. Electron Devices* **ED-56**, 1330 (2009)
29. Y.-T. Chen et al., *Appl. Phys. Lett.* **96**, 253502 (2010)
30. Y. Wang et al., *ECS Trans.* **33**, 487 (2010)
31. Y. Wang et al., *ECS Trans.* **41**, 243 (2011)
32. R. Chau, S. Datta, M. Doczy, B. Doyle, J. Kavalieros, M. Metz, High-k/Metal–gate stack and its MOSFET characteristics. *IEEE Electron Device Lett.* **25**(6), 408 (2004)
33. N. Nakagawa, H. Y. Hwang, and D. A. Muller, *Nat. Mater.* **5**, 204 (2006)
34. H. Watanabe, D. Matsushita, K. Muraoka, *IEEE Trans. Electron Devices* **53**, 1323 (2006)
35. S. Borowitz, *Fundamentals of Quantum Mechanics* (W. A. Benjamin, New York, 1967)
36. E. Merzbacher, *Quantum Mechanics* (Wiley, New York, 1961)
37. T. Ando, A. B. Fowler, F. Stern, *Rev. Mod. Phys.* **54**, 437 (1982)
38. H. Lueth, *Surfaces and Interfaces in Solid Materials* (Springer, Berlin, 1995)
39. C. Kittel, *Introduction to Solid State Physics* (Wiley, New York, 1967)
40. J.R. Hauser, K. Ahmed, Characterization of ultra-thin oxides using electrical CV and IV measurements, in *International Conference on Characterization and Metrology for ULSI Technology Proceedings* (1998) pp. 235–239
41. S. Krishnamurthy, S. Jallepalli, C.-F. Yeap, K. Hasnat, A.F. Tasch, C.M. Maziar, A computationally efficient model for inversion layer quantization effects in deep submicron N-channel MOSFETs. *IEEE Trans. Electron Devices* **43**(1), 90–96 (1996)

42. L.F. Register, A.F. Tasch, S.K. Banerjee, Understanding the effects of wave function penetration on the inversion layer capacitance of NMOSFETs. *Electron Device Lett.* **22**(3), 145–147 (2001)
43. V. Heine, Theory of surface state. *Phys. Rev.* **138**, A1689–A1696 (1965). MIGS
44. A. Rose, *Concepts in Photoconductivity and Allied Problems* (Wiley Interscience, New York, 1963)
45. D. Muñoz Ramo, J.L. Gavartin, A.L. Shluger, G. Bersuker, *Phys. Rev. B* **75**, 205336 (2007)
46. I.E. Tamm, *Z. Physik* **76**, 849 (1932)
47. W. Shockley, *Phys. Rev.* **56**, 317 (1939)
48. G.J. Gerardi, E.H. Poindexter, P.J. Caplan, N.M. Johnson, Interface traps and Pb centers in oxidized (100) silicon wafers. *Appl. Phys. Lett.* **49**, 348 (1986)
49. A.H. Edwards, Theory of the P_b center at the $\langle 111 \rangle$ Si/SiO₂ interface. *Phys. Rev. B* **36**, 9638 (1987)
50. J. Dong, D.A. Drabold, Atomistic structure of band-tail states in amorphous silicon. *Phys. Rev. Lett.* **80**(9), 1928–1931 (1998)
51. S. Kar, W.E. Dahlke, Interface states in MOS structures with 20–40 Å-thick SiO₂ films on non-degenerate Si. *Solid-State Electron.* **15**, 221–232 (1972)
52. A.V. Kimmel, P.V. Sushko, A.L. Shluger, G. Bersuker, *ECS Trans.* **19**(2), 3 (2009)
53. A. Toriumi, K. Kita, *ECS Trans.* **19**(1), 243 (2009)
54. H. Jagannathan, V. Narayanan, S. Brown, *ECS-Trans.* **16**, 19 (2008)
55. J. Tersoff, Schottky barrier heights and the continuum of gap states. *Phys. Rev. Lett.* **52**, 465–468 (1984)
56. M.R. Visokay, J.J. Chambers, A.L.P. Rotondaro, A. Shanware, L. Colombo, *Appl. Phys. Lett.* **80**, 3183 (2002)
57. J.-H. Lee et al., in *2002 Symposium on VLSI Technology Digest of Technical Papers*, 2002
58. Y.H. Wu, M.Y. Yang, A. Chin, W.J. Chen, C.M. Kwei, *IEEE Electron Device Lett.* **21**, 341 (2000)
59. H. Harris, K. Choi, N. Mehta, A. Chandolu, N. Biswas, G. Kipshidze, S. Nikishin, S. Gangopadhyay, H. Temkin, *Appl. Phys. Lett.* **81**, 1065 (2002)
60. C.N. Berglund, *IEEE Trans. Electron Devices* **13**, 701 (1966)
61. L.M. Terman, *Solid-State Electron.* **5**, 285 (1962)
62. A. Ali, H. Madan, S. Koveshnikov, S. Datta, *ECS Trans.* **25**(6), 271 (2009)
63. M. Heyns et al., *ECS Trans.* **25**(6), 51 (2009)
64. S. Kar, C. Miramond, D. Vuillaume, Properties of electronic traps at silicon/1-octadecene interfaces. *Appl. Phys. Lett.* **78**, 1288 (2001)
65. J. McNutt, C.T. Sah, *J. Appl. Phys.* **46**, 3909 (1975)
66. J. Maserjian, G. Petersson, C. Svensson, *Solid-State Electron.* **17**, 335 (1974)
67. J. Maserjian, in *The Physics and Chemistry of SiO₂ and the Si/SiO₂ Interface*, ed. by C.R. Helms, B.E. Deal (Plenum Press, New York, 1988)
68. B. Ricco, P. Olivo, T.N. Nguyen, T.-S. Kuan, G. Ferriani, *IEEE Trans. Electron Devices* **35**, 432 (1988)
69. K. Ahmad, E. Ibok, G. Bains, D. Chi, B. Ogle, J.J. Wortman, J.R. Hauser, *IEEE Trans. Electron Devices* **47**, 1349 (2000)
70. J.S. Brugler, P.G.A. Jespers, *IEEE Trans. Electron Devices* **16**, 207 (1969)
71. G. Groeseneken, H.E. Maes, N. Beltran, R.F. deKeersmaecker, *IEEE Trans. Electron Devices* **31**, 42 (1984)
72. S. Kar, S. Rawat, *ECS Trans.* **16**(5), 111 (2008)
73. K. Shiraishi, K. Yamada, K. Torii, Y. Akasaka, K. Nakajima, M. Konno, T. Chikyo, H. Kitajima, T. Arikado, Y. Nara, *Thin Solid Films* **508**, 305 (2006)
74. T.C. Poon, H.C. Card, *J. Appl. Phys.* **51**, 6273 (1980)
75. S. Kar, S. Varma, *J. Appl. Phys.* **58**, 4256 (1985)
76. B. Hoeneisen, C.A. Mead, *Solid-State Electron.* **15**, 819 (1972)
77. S. Kar, Characterization of silicon MOS tunnel diodes. *IEDM Tech. Dig.* **79** (1976)

78. S. Kar, Two limiting thinnesses of the ultrathin gate oxides, in *Silicon Nitride and Silicon Dioxide Thin Insulating Films*, ed. by K.B. Sundaram, M.J. Deen, D. Landheer, W.D. Brown, D. Misra, M.D. Allendorf, R.E. Sah, Electrochem. Soc. Proc., vol PV-2001-7, **60** (2001)
79. R. Chau, B. Boyanov, B. Doyle, M. Doczy, S. Datta, S. Harelend, B. Jin, J. Kavalieros, M. Metz, Silicon nano-transistors for logic applications. *Physica E* **19**, 1 (2003)
80. D.A. Muller, T. Sorsch, S. Moccio, F.H. Baumann, K. Evans-Lutterodt, G. Timp, The electronic structure at the atomic scale of ultrathin gate oxides. *Nature* **399**, 758 (1999). 24 June
81. S. Kar, S. Varma, *J. Appl. Phys.* **54**, 1988 (1983)
82. R. Nieh, R. Choi, S. Gopalan, K. Onishi, C.S. Kang, H.-J. Cho, S. Krishnan, J.C. Lee, Evaluation of silicon surface nitridation effects on ultra-thin ZrO_2 gate dielectrics. *Appl. Phys. Lett.* **81**, 1663–1665 (2002)
83. C.H. Lee, J.J. Lee, W.P. Bai, S.H. Bae, J.H. Sim, X. Lei, R.D. Clark, Y. Harada, M. Niwa, D.L. Kwong, Self-aligned ultra thin HfO_2 CMOS transistors with high quality CVD TaN gate electrode, in *2002 Symposium on VLSI Technology Digest of Technical Papers*
84. Y.-S. Lin, R. Puthenkovilakam, J.P. Chang, Dielectric property and thermal stability of HfO_2 on silicon. *Appl. Phys. Lett.* **81**, 2041–2043 (2002)

High Permittivity Gate Dielectric Materials

Kar, S. (Ed.)

2013, XXXII, 489 p. 325 illus., 168 illus. in color.,

Hardcover

ISBN: 978-3-642-36534-8