1 Introduction

1.1 The historical CMOS scaling scenario

Complementary Metal Oxide Semiconductor (CMOS) technology is nowadays the backbone of the semiconductor industry worldwide and the enabler of the impressive number of electronic applications that continue to revolutionize our daily life. The pace of growth of CMOS technology in the last 40 years is clearly shown in the so-called Moore's plot (see Fig.1.1 [1]), reporting the historical trend in the number of transistors per chip, as well as in the trends of many other circuit performance metrics and economic indicators.

Key to the success of CMOS technology is the extraordinary scalability of the Metal Oxide Semiconductor Field Effect Transistor (MOSFET). The word *scaling* denotes the possibility, illustrated in Fig.1.2 and Table1.1, of fabricating functional devices with equally good or even improved performance metrics but smaller physical dimensions. The design of scaled transistors starting from an existing technology has been driven initially by simple similarity laws aimed to maintain essentially unaltered either the maximum internal electric field (hence, to a first approximation, the device reliability) or the supply voltage (hence the system integration capability) [2].

According to these two scaling strategies, defined in Table1.1, all the lateral (primarily the gate width, W, and length, L_G) and the vertical physical dimensions (the thickness of the gate dielectric, t_{ox}, and the junction depth, x_j) should decrease from one technology generation to the next by a factor α, thus yielding an increase of the number of transistors per unit chip area by a factor of α^2. In order to proportionally reduce the channel depletion depth, the doping concentration in the substrate should increase by no less than a factor α. The intrinsic switching delay $\tau = CV/I$ is consequently reduced by a factor ranging between α^{-1} and α^{-2} in the constant field and constant supply voltage scaling scenario, respectively.

The constant field and constant supply voltage scaling rules are derived from quite simple one-dimensional models of the MOSFET electrostatics. These models and the rules above became inadequate to the design of MOS transistors as the gate length (L_G) approached one micron, thus leading to development of more sophisticated criteria. As an example, Table1.1 reports the mixed scaling rules proposed in [3] to design 0.25 μm MOSFETs, where different reduction factors are introduced for the geometrical dimensions (α) and the voltages (λ).

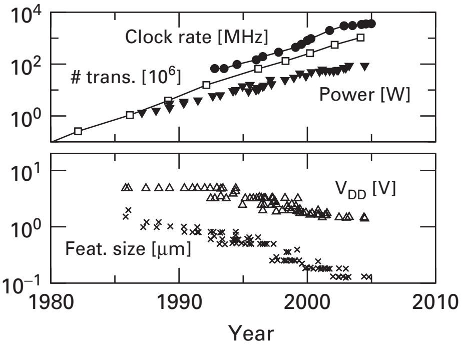

Figure 1.1 Progress in CMOS technology. Number of transistors in memory chips, clock rate, power supply voltage, power consumption, and minimum feature size.

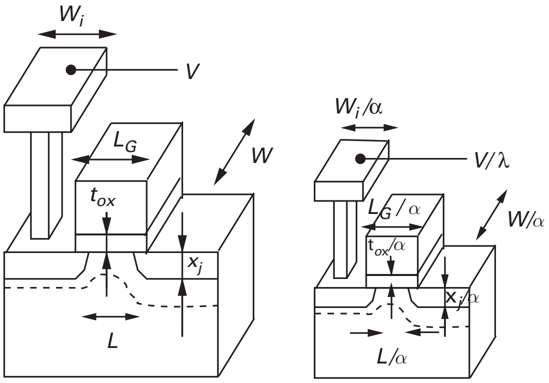

Figure 1.2 Bulk MOSFET scaling principles and corresponding scaling factors for geometrical dimensions (α) and voltages (λ). Note that L_G and L denote the gate length and the effective channel length, respectively.

Table 1.1 Scaling rules for CMOS technology. Note that α and λ denote the geometry and voltage scaling factors, respectively.

Parameter	Const.field scenario	Const.voltage scenario	Mixed scenario
Dimensions	$1/\alpha$	$1/\alpha$	$1/\alpha$
Voltages	$1/\alpha$	1	$1/\lambda$
Fields	1	α	α/λ
Doping	α	α^2	α^2/λ
Current	$1/\alpha$	α	α/λ^2
Capacitance	$1/\alpha$	$1/\alpha$	$1/\alpha$
Interconnect resistance	α	α	α
Switching delay	$1/\alpha$	$1/\alpha^2$	λ/α^2
Interconnect delay	1	1	1
Power delay product	$1/\alpha^3$	$1/\alpha$	$1/\alpha^2\lambda$
Power area-density	1	α^3	α^3/λ^3

Cambridge University Press
978-0-521-51684-6 - Nanoscale MOS Transistors: Semi-Classical Transport and Applications
David Esseni, Pierpaolo Palestri and Luca Selmi
Excerpt
More information

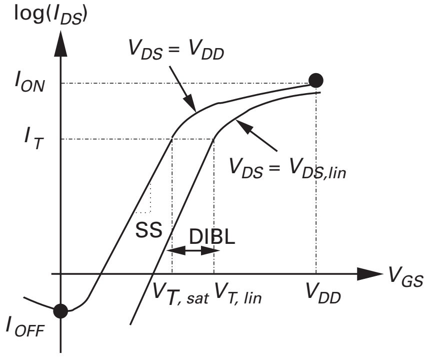

Figure 1.3 Definition of the main static performance metrics of a MOSFET. V_{DD} is the power supply voltage, I_T is a threshold drain current (typically 1 μA/μm). $I_{ON} = I_{DS}$ at $V_{GS} = V_{DS} = V_{DD}$; $I_{OFF} = I_{DS}$ at $V_{GS} = 0$ V and $V_{DS} = V_{DD}$; $V_{T,lin} = V_{GS}$ at $I_{DS} = I_T$ and $V_{DS} = V_{DS,lin}$; $V_{T,sat} = V_{GS}$ at $I_{DS} = I_T$ and $V_{DS} = V_{DD}$; subthreshold swing SS $= dV_{GS}/d[\log(I_{DS})]$; DIBL $= (V_{T,lin} - V_{T,sat})/(V_{DD} - V_{DS,lin})$.

In particular, since the thermal voltage K_BT/e, the band gap and the junction built-in voltage do not scale, the subthreshold swing (SS) of the transfer characteristic and the flatband voltage of poly-silicon gate MOSFETs remain almost invariant to scaling [4]. As a result, the two-dimensional distribution of the electrostatic potential inside the scaled device is distorted compared to that of the parent technology generation and so-called Short Channel Effects (SCE) become apparent as:

- a decrease of the linear and saturation threshold voltages ($V_{T,lin}$, $V_{T,sat}$, Fig.1.3) at short channel lengths, due to the penetration of the source and drain electric field lines in the channel region;
- a large sensitivity of the threshold voltage to the drain voltage (an effect denoted as DIBL, Drain Induced Barrier Lowering);
- an increase of the subthreshold swing SS.

Narrow channel effects, detrimental to control of the threshold voltage, also appear in the scaled technology.

An optimum choice of channel doping, junction depth and thickness of the gate dielectric is crucial to keep SCE under control. Accurate tailoring of the source and drain extensions below the spacers and reduction of parasitic source/drain resistances contribute as well to achieving good performance and high I_{ON}/I_{OFF} ratios. As a consequence of the increased complexity of this optimization task, during the eighties two- and three-dimensional CAD tools for numerical device simulation (mostly based on the Drift-Diffusion semiconductor device model [5–9]) have found widespread use in the semiconductor industry to assist process engineers in analysis and tuning of the doping profiles to counteract the short channel effects.

Starting from the early nineties, foresight studies on the scaling of CMOS technology have emerged from the joint efforts of associations such as the US Semiconductor

Industry Association (SIA) and later the International Technology Roadmap for Semiconductors (ITRS). The guideline documents on MOSFET scaling prepared by the ITRS [10] aim at the early identification of risk factors in the developments of the microelectronics industry, as well as at steering research toward the so called "red brick walls" which may impede further progress of this strategic technology.

In recent years, diversification of microelectronic applications has led to a differentiation of the ITRS for High Performance (HP), Low Power (LP) and Low STand-by Power (LSTP) applications [11]. Nevertheless, regardless of the specific market area, the semiconductor industry has steadily pursued the scaling of the device footprint, that is the area scaling, in spite of the increased complexity of the fabrication technology and growing fabrication costs.

To a different extent, all the roadmaps for the bulk MOSFET architecture nowadays share a common difficulty in finding the balance in the trade-off involving the containment of SCE (which demands high channel doping and gate dielectrics with small equivalent oxide thickness, EOT), the quest for high on-current (which requires high carrier mobility and low threshold voltage), and the need for low subthreshold leakage (which requires high threshold voltage, low subthreshold swing and relatively thick gate dielectrics). The performance metrics of the bulk MOSFET technology have steadily improved [12] but, as the minimum channel length entered the sub 0.1μm range, it became increasingly difficult to maintain the historical scaling trends by mere optimization of the conventional architecture. Due to complexity and cost, however, the introduction of significant innovations has always been deferred till the time when no real alternative was possible.

A prominent example in this respect is the replacement of SiO_2 (with its nearly ideal interface properties, large band gap, low trap density, etc.). In an effort to prolong the usability of the most popular dielectric in silicon microelectronics, nitrided SiO_2 layers (SiON) were adopted first [13–17], with undebatable advantages in terms of increased dielectric constant and beneficial effects against boron penetration in p-MOSFETs. It is only with the advent of 45 nm technology that the first breakthrough innovation at the heart of the bulk MOSFET architecture, namely the introduction of high-κ dielectrics, has started to become a reality [18–20].

Recently, the number of technology challenges putting at risk the scaling of the conventional bulk MOS transistor has increased. Fundamental studies suggest that the evolution of CMOS technology, as outlined in the ITRS, is leading the MOSFET to nearly achieve the ultimate performance expected for charge transfer switches [21–25]. However, it is also becoming clearer and clearer that significant innovations will be necessary to make the ultimate CMOS a reality.

Consistently, new options (the so called *technology boosters*) and new device concepts have been identified by the ITRS to flank the traditional dimension, doping and voltage scaling. These new options could give significant advantages in terms of intrinsic device performance, thus allowing microelectronics to maintain progress along the so called Moore's law. Recent developments in CMOS technology are thus outlining a *generalized scaling* scenario, which is briefly illustrated in the next section.

1.2 The generalized CMOS scaling scenario

For decades the basic architecture of the MOS transistor has not changed dramatically, although a large number of innovations, including new materials (e.g., new metals, low-κ dielectrics for interconnects, etc.) and new processes (e.g., shallow trench isolation, source/drain silicidation, lightly doped extensions, etc.), have been introduced to enable controlled device scaling to smaller dimensions. In recent years, however, CMOS scaling has become in a sense a definitely more diversified exercise.

To illustrate this point, Fig.1.4 shows a few of the advanced MOSFET architectures envisioned by the ITRS for future MOSFET scaling scenarios toward the ultimate limits.

In order to contain static power dissipation in the off state and guarantee the device reliability, gate leakage currents must be kept under control. The simultaneous need to increase the effective gate capacitance has led to exploration of the use of alternative gate insulators with a relative dielectric constant κ higher than that of SiO_2 and SiON [26, 27], which can provide a given equivalent oxide thickness (EOT) with a larger physical thickness with respect to SiO_2 and thus reduce the gate leakage. The introduction of metal gate electrodes (Fig.1.4.a) eliminates poly-silicon depletion, thus contributing increased capacitance, but generates Fermi level pinning issues [28, 29]. Unfortunately, almost all eligible high-κ materials degrade the channel mobility [26, 27, 30, 31] unless a thin SiO_2 interfacial layer is left above the channel, which conversely limits the increase of the gate capacitance. Completely new reliability problems are raised as well by the introduction of the high-κ insulators [32].

Reduction of the EOT is not enough to maintain a good electrostatic integrity, because the penetration of the drain field in the channel increases the DIBL and the subthreshold

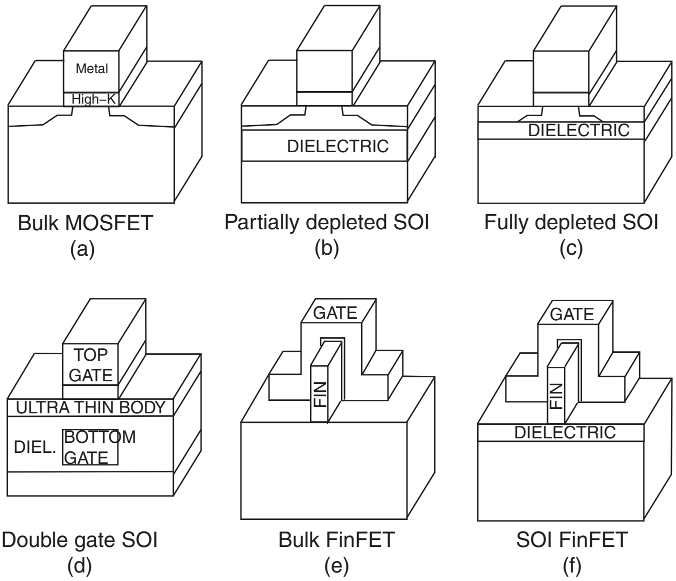

Figure 1.4 MOSFET architectures proposed for present and future CMOS technologies.

swing, unless the substrate doping is increased as well. In this respect, studies in the mid-nineties showed that improved control of the threshold voltage roll-off and low values of the subthreshold swing could be achieved at short channel lengths with ground plane architectures and, even better, with Silicon On Insulator (SOI) technologies [33]. Partially depleted SOI devices (PD-SOI, Fig.1.4.b) demonstrated some advantages over bulk MOSFETs, but the relatively large kink effect and the degradation of static and dynamic performance due to transient charge storage and self-heating effects impeded the blossoming of this technology. Moreover, SOI was and still is a costly technology option; except in a few cases, the portability of bulk designs to a SOI platform is not straightforward [34, 35].

The advent of the SIMOX and Unibond Smart-Cut processes [36] revitalized SOI as a credible technology option [37] and boosted research on high quality aggressively scaled SOI films [38–45]. For small enough silicon thickness the body of the transistor becomes fully depleted (FD-SOI, Fig.1.4.c); consequently, short channel effects, DIBL and subthreshold swing remarkably improve. The impact ionization induced kink effect disappears and good electrostatic integrity is achieved. The source/drain parasitic capacitance is also reduced because of the underlying buried oxide layer.

The SOI technology also facilitates the realization of double gate (DG, Fig.1.4.d) and gate all around (GAA) architectures that can bring CMOS even closer to its ultimate scaling limits by offering nearly optimum control of the gate over the channel [42, 46–48]. In fact, provided the film thickness is at least about 2.5 times the channel length, SCE are suppressed and nearly ideal subthreshold swing is observed (SS ≈ 60 mV/dec at room temperature) even in undoped channel transistors. Therefore a reduced fluctuation of V_T due to the discrete doping can be achieved but, at the same time, new means other than channel doping must be devised to tailor the V_T (e.g. workfunction engineering).

Another advantage of DG and GAA architectures is that, in the direction perpendicular to the transport, the average electric field at given inversion charge per channel is reduced compared to bulk devices because good electrostatic control can be achieved with essentially undoped films; hence, the carrier mobility is larger. Moreover, due to the double channel a DG device provides the same total inversion charge at lower effective field compared to single gate SOI; hence it can achieve the same I_{ON} of a single channel device at smaller gate voltages: a clear advantage in view of low voltage operation.

The FinFET technology (Figs.1.4.e and 1.4.f) provides an alternative approach to fabricating DG transistors [49]. In narrow FinFETs the conduction takes place mostly along sidewalls normal to the wafer plane and, in essence, a double gate device is obtained [50]. If the fin is large, instead, a significant fraction of the current flows along the top interface and the device is more appropriately referred to as a triple gate transistor.

The process complexity, variability, and cost of SOI and FinFET technology tend to offset the advantages offered in terms of scaling, thus leaving room for prolonged efforts on bulk MOSFET optimization. In particular, strained silicon technology and optimization of the crystal orientation are very effective means of boosting the mobility and I_{ON} of both n-MOS and p-MOS devices [51–57]. Indeed, the strain in the crystal lattice has a remarkable impact on the band structure, hence on the electrostatics and the transport properties of the device. With an appropriate combination of strain type, magnitude and orientation with respect to the crystal axes and the transport direction, on-current enhancements of up to 20–30% for sub-50 nm channel lengths have been demonstrated [58–60]. The remarkable success of strained silicon technology is keeping bulk MOSFET architecture competitive; as a result, the year of expected introduction of advanced SOI technology options has recently been postponed by the ITRS [61, 62].

To improve the device performance further it has also been proposed to replace the silicon channel with alternative semiconductors characterized by enhanced transport properties. As an example, bulk III-V materials are known to have superior electron mobility with respect to silicon, whereas hole mobility is high in bulk germanium. These considerations have led to a search for new ways to locally grow islands of different semiconductors on silicon substrates [63–65] and to develop compatible high quality gate stacks [66–70]. Studies have flourished aimed at assessing if alternative channel materials can bring real advantages in terms of inversion layer mobility and overall device performance [63, 71–75].

Last but not least, we emphasize that extrinsic parasitic components (source/drain resistances and overlap capacitances) may jeopardize the advantage of having smaller and faster intrinsic transistors. This is especially true for FinFETs and ultra-thin body fully depleted SOI MOSFETs, where the limited SOI film thickness implies a high series resistance. Elevated source/drain technology and non-overlapped devices alleviate these issues [76–83]. To boost the device performance even further, metallic source and drain technology has been proposed. By exploiting doping segregation, a pile-up of the dopants at the metal–semiconductor junction is obtained which relieves the detrimental effects of Schottky barrier formation [84]. Careful selection of the metal can possibly lead to achieving high current drive [85]. Variability due to fluctuations of the tail of dopants in the channel is also expected to decrease thanks to these technology improvements.

1.3 Support of modeling to nano-scale MOSFET design

As illustrated in the previous section, new materials and device architectures are expanding the design space to be explored for future CMOS and nano-electronic technologies. Single gate SOI, double gate SOI, FinFET, MuGFET, gate all around and nanowire device architectures are being investigated as possible successors of the conventional planar bulk MOSFET [86]. Gate metal workfunction, silicon body thickness, stress–strain distribution, gate stack composition, source, drain and channel material are only a few of the additional variables that it is necessary to engineer for the existing and future MOSFET generations.

The design and optimization of nano-transistors exploiting these new options demand general purpose models to describe electrostatic and transport phenomena at the nanoscale in an unprecedented variety of materials, with a reasonably predictable degree of accuracy and with affordable computation time. The established Drift-Diffusion model available in conventional TCAD tools is presently inadequate for the purpose.

In this respect, it is important to consider the substantial quantum mechanical effects in the direction perpendicular to the transport plane which are emphasized by size induced confinement in ultra-thin body architectures with silicon thickness below 10 nm. Carrier quantization decreases the effective gate capacitance (due to the combined effects of finite inversion layer thickness and dead spaces at the SiO_2 interfaces [87–89]) and reduces the inversion charge for a given gate voltage, thus altering the threshold voltage. The appearance of subbands affects the carriers' scattering as well, with remarkable implications for both the low and the high field transport characteristics of the inversion layer.

Quantum confinement is especially strong at the top of the potential energy barrier that governs carrier injection from the source to the channel region (the so called *virtual source*, [90, 91]). Since high levels of charge are desired in the on-state, the carrier gas becomes highly degenerate and the average carrier velocity becomes gate bias dependent.

Another relevant aspect concerning transport is that when the gate length L_G scales below a few tens of nanometers the mean free path in the channel is expected to become comparable to the device length [92, 93]. The fraction of the carrier population that reaches the drain without suffering scattering events tends to increase and the effects related to far from equilibrium transport become important. However, even if rare, scattering events in the channel cannot be neglected, because they affect the carrier density and thus the potential profile along the channel and contribute to shaping the potential energy barrier at the source and to setting the I_{ON} [94]. A sound description of transport in MOSFETs should cover the transition between conventional drift-diffusion and purely ballistic transport, and should obviously include all the relevant scattering mechanisms, especially those related to the introduction of new dielectric or semiconducting thin films.

Tunneling through the source barrier and band-to-band tunneling at the drain end of the channel may also become relevant, especially in the low band gap, small tunneling mass semiconductors being considered for ultimate CMOS [95–100]. Degraded I_{OFF} and subthreshold swing SS are expected if these leakage mechanisms are not kept under control.

The design and optimization of future nano-transistors require us to understand and master all these physical effects and their interrelations in an increasingly large number of materials and device architectures. A broad matrix of combinations must be evaluated and the device simulation can considerably facilitate this process, provided that predictive models are available to reduce the risk and cost of fabrication trials and errors.

Historically the attention of the industry toward the field of modeling and simulation has been mostly driven by the need to steer the selection of process and device parameters for incremental improvements of existing technologies. The broad spectrum of present day scaling scenarios has raised new interests in device modeling and simulation. New theories and new models to describe the links between the band structure of the materials, the device electrostatics, the transport and the performance have become of utmost importance. This new perspective is well expressed by the ITRS roadmap

[10], which devotes a full chapter to modeling and simulation and reiterates the quest for renewed efforts in the modeling of MOSFETs incorporating all the technology boosters of interest.

In this respect it is worth noting that the band gap, the density of states, the carriers' mobility and the other physical properties of the thin, possibly strained semiconductor layers used in fully depleted single or double gate SOI and FinFETs cannot be simply extrapolated from the corresponding properties of the bulk material. The widespread exploitation of stress and strain, and the possible use of alternative channel materials (germanium, silicon–germanium alloys, gallium arsenide) demand models to describe the subband structure and the transport parameters of quantized inversion layers (group velocity, effective mass, scattering rates, mobility, etc.) for both electrons and holes. These models should be general enough to tackle various substrate crystal orientations with respect to the quantization and the transport directions, and accurate enough to predict the stress-strain, film thickness and bias dependencies.

It is clear then, that exploring by simulation the design space of new nano-scale CMOS transistors demands a large innovative effort in physically based and in TCAD oriented modeling, which for decades has been mainly focused on unstrained silicon transistors fabricated almost exclusively on (001) wafers. Physically sound, modular and robust device modeling frameworks are necessary, where new physical effects can be added and related to the device performance, possibly starting from the physical properties of new materials. These frameworks should be general enough to include quantization effects on both electrostatics and carrier transport and to encompass all conduction regimes from drift-diffusion to fully ballistic.

1.4 An overview of subsequent chapters

Stimulated by recent developments in nano-electronics and inspired by the scenario outlined in the previous sections, we wrote this book to describe in detail the semi-classical modeling of carrier transport in modern nanoscale MOSFETs, accounting for the significant quantization effects that enforce the formation of electronic subbands in the transistors inversion layer. In particular, in the framework of this semi-classical model, the Schrödinger equation is used to calculate the quantum energy levels and the wave-functions of the inversion layer while a system of coupled Boltzmann transport equations describes the transport in the subbands. The Poisson equation is solved iteratively with the Schrödinger and the Boltzmann equations until convergence is reached to a fully self-consistent solution of the whole electrostatic and transport problems. Moreover, we illustrate a relevant implementation of the model, which we concisely denote as *multi-subband Monte Carlo* because it relies on use of the Monte Carlo method to solve the Boltzmann equations in the inversion layer subbands.

We have enhanced the book with a broad set of simulation results mostly obtained with the multi-subband Monte Carlo implementation of the model. These were selected to illustrate in detail how the physical elements of the semi-classical transport model in inversion layers affect the operation of modern MOSFETs.

With these objectives in mind, the book begins by recalling in Chapter 2 the elements of the semi-classical treatment of carrier transport in bulk crystals. In particular, we introduce the fundamental results regarding electrons in periodic crystalline lattices and the band structure of bulk crystals. We then describe a few methodologies to compute the conduction and valence band structure in bulk semiconductors and the simplest analytical approximations commonly used to model the dispersion relation in the proximity of the band edges. The last paragraphs of the chapter introduce the foundations of the semi-classical model of carrier transport by using a wave-packet representation of the electrons. We derive the semi-classical equations of motion under the action of slowly varying potentials and introduce the Fermi golden rule for the treatment of carrier scattering due to the action of rapidly fluctuating potentials.

Chapter 3 develops the effective mass approximation and the $\mathbf{k}\cdot\mathbf{p}$ quantization models for, respectively, electron and hole inversion layers. A full band quantization model based on the linear combination of bulk bands method is described as well, since it can serve as a useful reference to check the validity of the simpler quantization models in conditions of strong confinement, such as those present in ultra-thin semiconductor films. From there, the chapter moves to the calculation of carrier densities accounting for the density of states in a two-dimensional carrier gas and finally to self-consistent solutions of the Poisson and Schrödinger equations.

Chapter 4 contains an extensive theoretical treatment of scattering for carriers in inversion layers. Starting from the envelope eigenfunctions and eigenvalues and exploiting the Fermi golden rule, Coulomb, surface roughness, and phonon scattering mechanisms are analyzed in detail for both electrons and holes. The static and dynamic screening of the scattering potential produced by the inversion layer charge is also addressed. We have tried to provide a clear and pedagogical presentation of these topics. Particular attention was devoted to justifying and discussing the approximations behind the mathematical developments.

After Chapters 3 and 4, which provide the quantum mechanical foundations for the treatment of the two-dimensional carrier gas, we continue with Chapter 5 aimed at a description of the set of coupled BTEs for the subbands in the inversion layer. The case of free electrons and holes is treated first, to underline the connections to the semi-classical transport concepts explained in Chapter 2. Several examples clarify the expression of the driving force for carriers' motion in cases of practical relevance.

Chapter 5 describes also the solution of the BTE in inversion layers by means of the widely used Momentum Relaxation Time (MRT) approximation, whose usefulness and validity limits are discussed. The recently proposed ballistic and quasi-ballistic MOSFET models are then derived from the solution of the BTE where the terms related to scattering are neglected. These derivations allow us to clarify the approximations behind these popular models and are instrumental in introducing many concepts useful for interpretation of numerical simulations.

The discussion of solution methods for the BTE continues with Chapter 6, which is devoted to the Monte Carlo method. Here again the free carrier gas is treated first, but the multi-subband case is also specifically addressed at the end of the chapter. Many non-trivial technical details arising in the practical implementation of the method