Raghu Murali *Editor*

Graphene Nanoelectronics

From Materials to Circuits

Springer

Graphene Nanoelectronics

Raghu Murali
Editor

Graphene Nanoelectronics

From Materials to Circuits

Foreword by Jeff Welser

Springer

Editor
Raghu Murali
Nanotechnology Research Center
Georgia Institute of Technology
Atlanta, GA, USA

ISBN 978-1-4614-0547-4 e-ISBN 978-1-4614-0548-1
DOI 10.1007/978-1-4614-0548-1
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011946064

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

For almost a half century, the semiconductor industry has managed to double the number of transistors on an integrated circuit chip every 18–24 months with such consistency that it has been dubbed Moore's Law, after Gordon Moore who first noted the trend in 1965. [1] For the past 30 years, the primary technology for the industry has been the silicon complementary metal-oxide-semiconductor (CMOS) field-effect transistor (FET), and it has been the ability to shrink this device according to Robert Dennard's scaling theory [2] that has largely enabled Moore's Law to continue for so long. The result has been exponentially increasing performance per dollar in both integrated circuits (ICs) and all of the electronic and information technology (IT) systems they enable. Not only has scaling resulted in growing worldwide semiconductor industry revenues from $20B in 1980 to almost $300B in 2010, but it has also played a major role in driving the overall economy: in the United States, it has been estimated that from 1995 to 2005, while IT industries only made up 3% of Gross Domestic Product (GDP), they accounted for 25% of overall economic growth. Taken as a whole, "these industries contribute more to economy-wide productivity growth than all other industries combined." [3]

However in the early 2000s, as the gate length for the Si FET crossed into the sub-100 nm range and the traditional gate insulator approached 10 nm thicknesses, the industry's ability to continue to follow Dennard's rules for decreasing dimensions and supply voltage faltered. Subsequent technology generations have seen exponentially increasing active switching and passive leakage powers, limiting the ability to take full advantage from scaling devices and doubling transistor counts. The result has been an increased focus on new materials and/or devices that could continue the historical scaling trends.

Since its re-discovery through the ingenious use of scotch tape by Novoselov, Geim and co-workers [4] in 2004, graphene has emerged as an exciting new material, with potential to impact many areas of technology. The number of publications on graphene has exploded over the past 5 years, with new theoretical claims and experimental breakthroughs seeming to happen almost monthly. However, it is a long journey from those eureka moments to finding real

applications for graphene in new devices and technology. This book attempts to take a comprehensive look at graphene in the context of nanoelectronic applications, with an eye towards highlighting the key physics, material properties, and fabrication challenges that would most interest a researcher looking to apply it to new transistors and IC technologies. It is hoped this overview will be useful for both introducing the field to new investigators, as well as offering an organized reference to those already immersed in the field.

To set the stage for understanding the challenges ahead for any new nanoelectronics technology, it is useful to step back and consider critically the status of the current technology. In the first chapter, "CMOS Performance Scaling," the physics of scaling a Si FET down to its current gate length of just tens of nanometers and beyond are examined, highlighting the challenges of maintaining performance while managing both active and passive power. The prospect of continuing to add new innovations, such as strain and new materials for both the gate and channel regions of the device, are also discussed to understand the potential paths for extending FET scaling to its limits.

In the next two chapters, the prospects for graphene in helping to address some of these challenges in future FETs are considered. "Electronic Transport in Graphene" gives a thorough overview of the basic graphene material properties and physics, with a specific eye on how these properties can be exploited for electronic transport devices. Building on this base, "Graphene Transistors" considers how to apply the material specifically to transistors, both for digital and analog applications. This chapter thoroughly covers all of the considerations that go into making a good transistor, and while the high carrier velocity in graphene makes it attractive for many applications, the lack of a natural bandgap in the unpatterned material adds new challenges, particularly for digital switches.

Going beyond simply improving existing device structures, the "Alternative State Variables for Graphene Transistors" chapter considers entirely new ways of exploiting graphene's unique physics. Rather than simply building transistors that modulate current flow to manipulate data, this chapter looks at the possibility of representing information with other state variables, such as spin, pseudospin, or even mechanical motion. In each case, graphene offers new ways to manipulate these novel states that may open the door to a whole new class of devices. However, to build these devices into circuits may require an even bigger change in the "wiring" between them. "Transport of Novel State Variables" considers the physics of moving various state variables—from electronic charge to carrier spin to plasmons—between devices in graphene, and compares the potential performance and energy dissipation of these interconnects with state-of-the-art CMOS. Indeed, transport may end up being the critical factor in choosing future state variables and devices.

While graphene offers many exciting new technology opportunities, they all depend on the ability not only to successfully produce high-quality graphene films, with single or few atomic layer thicknesses, but also to pattern these films and integrate them with other materials to create full device structures. The last portion of this book considers several aspects of these challenges, starting with three

chapters on graphene film growth: "Formation of Epitaxial Graphene," "Graphene Growth by CVD Methods," and "Chemical Approaches to Produce Graphene Oxide and Related Materials." While all of these approaches offer significant advantage over the initial "scotch tape method" of graphene film formation, all face clear challenges in producing uniform, low-defect layers, particularly over the large areas necessary for making a manufacturable technology. Finally, given the importance of gate structures in almost all devices, the last chapter "Atomic Layer Deposition of Dielectrics on Graphene" looks at methods for depositing insulators with high-quality interfaces on the thin graphene films.

The early days of research into any new material are often filled with a mix of hype and pessimism: while some are quick to exclaim the revolutionary potential for creating whole new technology paradigms, others are equally eager to point out the impossibility of ever supplanting the existing infrastructure. Hopefully this book can contribute by offering a balanced view of both the opportunities and challenges presented by graphene, and encourage continued, methodical work on how best to utilize this exciting material in future nanoelectronic systems.

Jeff Welser
Director, Nanoelectronics Research Initiative (NRI)
Semiconductor Research Corporation (SRC) &
IBM Almaden Research Center, San Jose, CA, United States

References

1. G. E. Moore, "Cramming more components onto integrated circuits," Electronics, vol. 38, no. 8, pp. 114–117, 1965.
2. R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design for ion-implanted MOSFET's with very small physical dimensions," IEEE J. Solid-State Circuits, vol. SC-9, no. 5, pp. 256–268, Oct. 1974.
3. D. Jorgenson, "Moore's law and the emergence of the new economy," Semiconductor Industry Association, Washington, DC, 2005 Annual Report, 2005.
4. K.S. Novoselov, A.K. Geim, S.V. Morozov, D. Jiang, Y. Zhang, S.V. Dubonos, I.V. Grigorieva, and A.A. Firsov, "Electric field effect in atomically thin carbon films," Science, vol. 306, pp. 666–669, 2004.

Contents

Contributors

Dimitri A. Antoniadis Microsystems Technology Laboratories, Massachusetts Institute of Technology, Cambridge, MA, USA

Kosmas Galatsis University of California, Los Angeles, CA, USA

Nelson Y. Garces U.S. Naval Research Laboratory, Washington, DC, USA

D. Kurt Gaskill U.S. Naval Research Laboratory, Washington, DC, USA

Ajey P. Jacob Intel Corp., Portland, OR, USA

Ali Khakifirooz IBM Research, San Jose, CA, USA

Jing Kong Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA

Raghu Murali Nanotechnology Research Center, Georgia Institute of Technology, Atlanta, GA, USA

Azad Naeemi School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Luke O. Nyakiti U.S. Naval Research Laboratory, Washington, DC, USA

Shaloo Rakheja School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Alfonso Reina Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA

Alexander Shailos University of California, Los Angeles, CA, USA

Alexander Sinitskii Department of Chemistry, University of Nebraska – Lincoln, Lincoln, NE, USA

James M. Tour Department of Chemistry, Rice University, Houston, TX, USA

Kang L. Wang University of California, Los Angeles, CA, USA

Virginia D. Wheeler U.S. Naval Research Laboratory, Washington, DC, USA

Jun Zhu Department of Physics, Penn State University, University Park, PA, USA

Chapter 1
CMOS Performance Scaling

Ali Khakifirooz and Dimitri A. Antoniadis

CMOS transistor density has been scaled exponentially in the past two decades and the intrinsic device performance has followed a commensurate scaling trend. Prior to the 90-nm node, mere shrinking of the device dimensions, following Dennard's scaling theory, was sufficient to guarantee increased device performance; beyond the 90-nm node, new innovations were necessary to continue the historical performance scaling trend. Strain engineering and high-k/metal gate technologies were the two major innovations that made the commensurate performance scaling in the past decade possible. However, it appears that new device structures and performance boosters will continue to be the need of the future. This chapter provides a basic overview of MOSFET scaling trend, followed by a discussion of MOSFET operation in deca-nanometer scale based on the so-called virtual source injection model. A simple analytical model for transistor *I–V* characteristics and intrinsic transistor delay is provided and used to quantify the historical trends of MOSFET performance scaling. Carrier velocity is shown to be the main driver for the continued MOSFET performance increase. Finally, the prospect of velocity increase is reviewed for strained Si, Ge, and compound semiconductors.

1.1 Introduction

The integrated circuit industry has witnessed a tremendous growth in the past four decades. This has been enabled by continuous scaling of transistors as described by Moore's law, which states that the number of the transistor per chip doubles every

A. Khakifirooz (✉)
IBM Research, San Jose, CA, USA
e-mail: khaki@us.ibm.com

D.A. Antoniadis
Microsystems Technology Laboratories, Massachusetts Institute of Technology, Cambridge, MA, USA

R. Murali (ed.), *Graphene Nanoelectronics: From Materials to Circuits*,
DOI 10.1007/978-1-4614-0548-1_1,

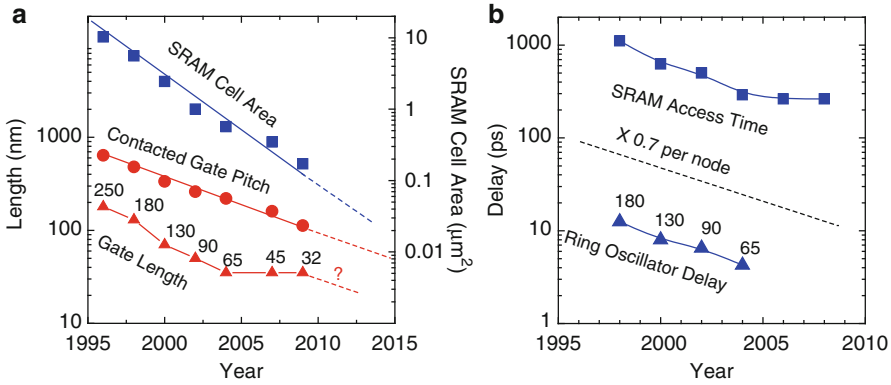

Fig. 1.1 Scaling trend of MOSFET geometry over the past 15 years (**a**) and the associated decrease in delay (**b**). MOSFET pitch is scaled by a factor of 0.7 per technology node to accommodate for doubling of transistor count. Up to the 65-nm node, circuit delay has followed the same pace of scaling by a factor of 0.7 per node (Data from Intel)

2 years. As a result, the chip area and thus the cost of delivering a given functionality has dropped exponentially during this period. What happened in practice, however, was that chip area remained almost constant and the chip functionality increased.

The key enabler for the exponential scaling of the CMOS technology is the scaling of the MOSFET pitch, which has been historically scaled by a factor of 0.7 per technology generation as illustrated in Fig. 1.1. Other transistor dimensions have been scaled with almost the same pace, according to Dennard's scaling theory [1]. Note that technology scaling does not necessarily require proportional scaling of the MOSFET gate length as long as it fits the pitch. In fact, Fig. 1.1 shows that starting from 180-nm technology, the gate length was shrunk faster that the pitch, perhaps to achieve higher speed. However, gate length scaling has virtually stopped since the 65-nm node. Of course, this trend cannot continue forever, and at some point, most likely at the 22-nm node, the gate length needs to be scaled down. Up to the 65-nm node, transistor delay followed the pace of 30% reduction per node. This was achieved by mere dimensional scaling down to the 130-nm node and then through various strain engineering techniques since the 90-nm node. However, it appears that delay scaling has slowed down since the 45-nm node. This is due to the fact that parasitic components associated with a MOSFET do not follow the scaling trend and their relative importance increases as transistors shrink.

Figure 1.2 shows the main parasitic components associated with a state-of-the-art MOSFET. The effective fringing capacitor, C_f^*, which consists of the inner fringe, C_{if}, outer fringe, C_{of}, and overlap capacitance, C_{ov}, is approximately 0.5 fF/μm and does not scale with gate length [3]. In fact, due to the proximity of source/drain contacts to the gate electrode, future technology nodes will suffer from a larger parasitic capacitance [4]. On the other hand, the source/drain series resistance, which consists of the silicide/semiconductor contact resistance and the resistance associated with the heavily doped S/D regions and the extensions, is about

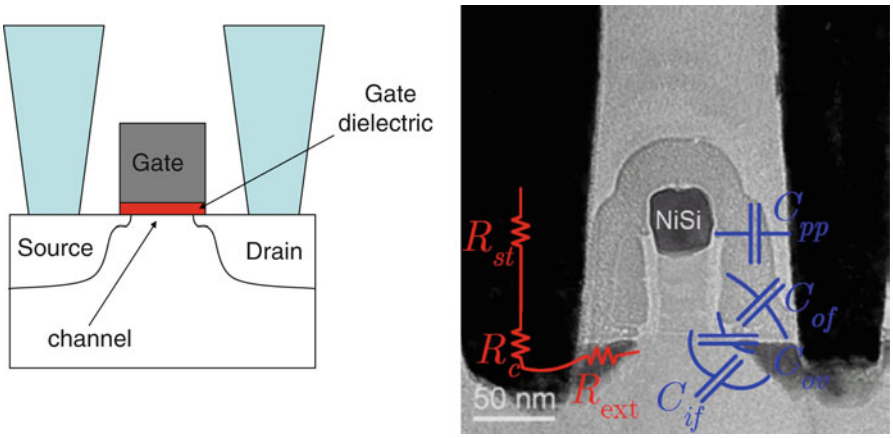

Fig. 1.2 MOSFET structure and its main parasitic components: the effective fringing capacitance, C_f^*, which consists of the inner fringe, C_{if}, outer fringe, C_{of}, and overlap capacitance, C_{ov}, is approximately 0.5 fF/μm and does not scale with the gate length. In fact, due to the proximity of the gate electrode and S/D contacts, another term, C_{pp}, is added as the device pitch is further scaled down. The source/drain series resistance consists of the silicide/semiconductor contact resistance, R_c, and the series resistance associated with the heavily doped S/D regions and extensions, R_{ext}, and does not scale down very well. As these devices are further shrunk, an additional series resistance due to finite conductivity of the contact studs, R_{st}, is added to the total series resistance (TEM image is reprinted with permission from [2] © 2005 IEEE)

80 Ω μm per side for modern NFETs and does not scale very well. Furthermore, as the devices are scaled, the resistance associated with finite conductivity of the metal contacts is becoming more significant and remedies, such as Cu contacts, were introduced to minimize this additional component.

An equally important issue for state-of-the-art CMOS is the scaling of the supply voltage (V_{DD}) to keep the power consumption under control as well as to maintain the device reliability. While the supply and threshold voltage (V_T) followed a similar trend in the earlier technology nodes, threshold voltage scaling was eventually slowed down to control the exponentially growing standby power. This means that less gate over drive ($V_{DD} - V_T$) is available as the devices are scaled. In conjunction with the increased importance of the parasitic components, this has already decelerated delay scaling.

A possible solution to compensate for the performance drop imposed by the loss of gate overdrive and the increase in the parasitic components is to improve transport properties of the channel by employing new materials. In fact, over the past decade, strain engineering has been extensively used to alter the transport properties of the Si channel and thereby continue the historical trend of the performance scaling. However, it appears that strained silicon is approaching its limit while the effectiveness of some of the strain engineering methods diminishes as device pitch is scaled below 100 nm. New channel materials are being studied to enable continued transistor performance scaling in the future.

1.2 Basics of MOSFET Operation

As shown in Fig. 1.2, the MOSFET is a three-terminal device, where the gate controls the conductivity of the channel region and thus flow of current between source and drain. The source and drain are doped with a high density of impurity atoms and with a polarity opposite to that in the channel region. For example, in a silicon n-MOSFET the source and drain are doped with arsenic or phosphorus with a density above 1×10^{20} cm^{-3}, while the channel region is doped with boron or indium with a density less than 1×10^{19} cm^{-3}. The gate is separated from the channel by a dielectric with a thickness of about 1 nm. At low gate voltages, there is no electron conduction between the source and drain. When the gate voltage is increased above a threshold voltage,[1] a thin layer of electrons called inversion layer is formed in the channel region, next to the gate dielectric, and makes the flow of electrons from source to the drain possible. In that sense a MOSFET can be considered as a switch; when the gate voltage is smaller than a threshold voltage, there is limited current flow and the transistor is "off", and when the gate voltage is above the threshold voltage the transistor is "on".

Of course, the MOSFET is not an ideal switch. When the transistor is "off", the source-drain current flow is not zero, instead it depends exponentially on the gate voltage as seen in Fig. 1.4. The voltage spent to drop the current by one order of magnitude is called subthreshold swing; ideally 60 mV/dec at room temperature, but typically 80–100 mV/dec for state-of-the-art transistors. Furthermore, unlike an ideal switch, the threshold voltage depends on the voltage across the switch, i.e., the drain to source voltage. This phenomenon is known as drain-induced barrier lowering (DIBL) as will be discussed further in the next section. Ideally, DIBL should be zero but it is typically 50–200 mV/V in state-of-the-art MOSFETs.

To minimize the subthreshold swing and DIBL, the gate needs to have a stronger coupling to the channel than the drain. This is achieved by aggressive scaling of the gate dielectric thickness and by making the source/drain extensions shallow. Also, it will help to confine the carriers closer to the gate dielectric by either placing a heavily doped well or halo (p-type for an NFET) as done in bulk MOSFETs or by making the substrate only a few nanometers thick as done in fully-depleted silicon-on-insulator (FDSOI) devices.

[1] Note that the definition of the threshold voltage is somewhat arbitrary as the transition from off-state to on-state, also known as strong inversion, is gradual. The transition region is often called weak inversion. Several definitions for the threshold voltage are given in the literature. The two most common definitions that are based on the *I–V* characteristics of the transistor are (1) constant current threshold voltage, where V_T is defined as the gate voltage at which the drain current is equal to an empirically-defined current, usually around 10^{-7} A/*WL*, with *W* and *L* being the gate width and length in micrometer, respectively, and (2) extrapolated threshold voltage, obtained by drawing the tangent to the *I–V* curve at the point where the transconductance is maximum and finding the intercept with the *x*-axis.

Obviously, as a switch, it is desired that a MOSFET delivers a high I_{on}/I_{off} ratio. In a typical CMOS logic circuit most of the transistors are in the off state at any given time. Reducing the off-current thus minimizes the static power that is wasted even when the circuit is in the idle mode:

$$P_{\text{Static}} = \sum I_{\text{off}} V_{DD}$$

On the other hand, since the transistors in the logic circuit mostly drive capacitive loads, higher on-current enables faster switching:

$$\text{delay} \sim C_{\text{eff}} V_{DD} / I_{\text{on}}$$

A more accurate delay metric is introduced in Sect. 1.5.

1.3 Physics of MOSFET Operation in Deca-Nanometer Scale

The essential physics of state-of-the-art MOSFET operation can be captured by the channel potential barrier model. Note that this is a descriptive model and the actual shape of the barrier depends on the details of the device structure, bias conditions, semiconductor band structure, and carrier transport.

A potential barrier in the channel controls the flow of carriers from source to drain. In the subthreshold regime, i.e., gate voltage smaller than threshold voltage, there is a large barrier to the flow of carriers from source to drain, Fig. 1.3a. The barrier height is linearly controlled by the gate voltage but the thermionic emission of carriers over the barrier is exponentially proportional to the barrier height and hence the current is exponentially proportional to the gate voltage and the temperature. Of course, if the channel is short enough, the barrier will be so narrow that some carriers can directly tunnel from the source to the drain, but this is not the case for a well-designed MOSFET. Increasing the drain voltage in the subthreshold regime slightly decreases the barrier height and makes the barrier slightly narrower, Fig. 1.3b. This is called DIBL and is usually modeled as a reduction in the threshold voltage proportional to the drain voltage. However, it should be noted that the barrier height reduction is not a linear function of the drain voltage and is usually stronger at smaller drain voltages.

When V_G is above the threshold voltage, the barrier is lowered enough so that many carriers can flow over the barrier and towards the drain. The "inversion" charge density in the channel is determined by a Fermi distribution that depends on the difference between the Fermi level and the local potential. In the near equilibrium case shown in Fig. 1.3c, the barrier is extended over almost the entire channel. Therefore, carrier transport depends on the scattering rate over the entire channel. In this case, the transistor current depends linearly on the drain voltage and hence the name linear regime. When the drain voltage is high enough, as in Fig. 1.3d, the

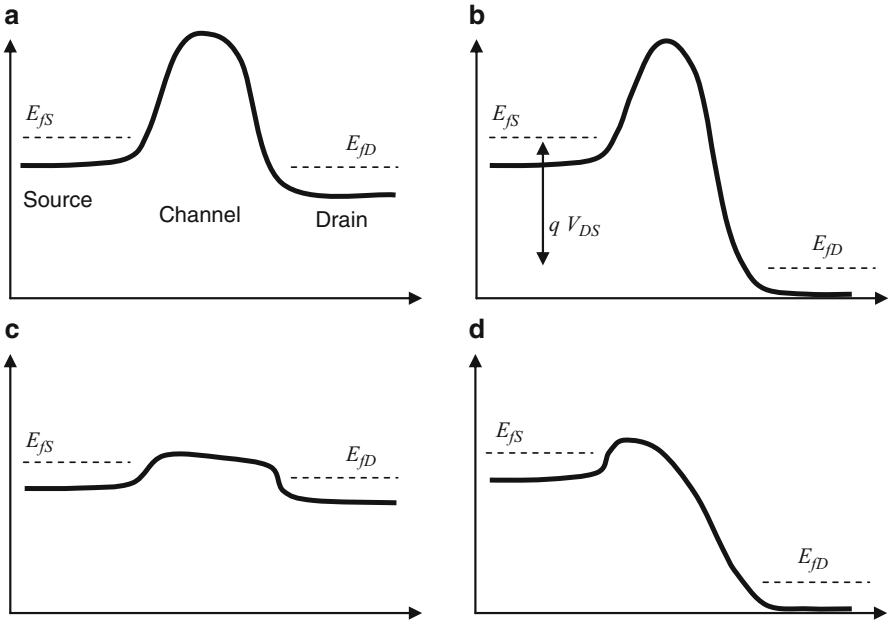

Fig. 1.3 Channel potential barrier at different bias conditions (**a**) in the subthreshold regime, i.e. $V_G < V_T$, there is a large barrier impeding the flow of carriers from the source to drain. (**b**) The barrier is lowered and slightly narrowed as the drain voltage is increased. The effect is commonly known as the drain-induced barrier lowering (DIBL) and is modeled as a reduction in the threshold voltage proportional to the drain voltage. With V_G above the threshold voltage, the barrier is low enough that many carriers can go over the barrier and flow to the drain. (**c**) With a small drain voltage, such a small barrier extends over the entire channel, so the current depends on the scattering mechanism over the entire channel. This is the so-called linear regime, where the current depends linearly on the drain voltage. (**d**) If the drain voltage is high enough, the barrier extends only over a small region near the source. Carrier transport is effectively controlled by the rate they overcome the barrier at the top of the barrier, the so-called virtual source

barrier extends only over a small region near the source. Carrier transport depends on the rate that they overcome the barrier at its peak, the so-called virtual source. Once a carrier goes past this point, it is unlikely that future scattering events cause the carrier to go back to the source. Although, in its simplest form it is assumed that the transistor current only depends on the scattering rates at the vicinity of the virtual source, it should be noted that the exact shape of the barrier depends on carrier density, which in turn depends on carrier distribution along the channel according to Fermi statistics. To satisfy charge continuity along the channel, the charge density in turn depends on the average carrier velocity, which depends on scattering rates over the entire channel. It is also assumed that in this case the transistor current is almost independent of the drain voltage, except for the reduction of threshold voltage due to DIBL. In other words, the transistor current saturates as drain voltage increases and hence the name saturation regime.

1.4 Simple MOSFET Model

In this section, a simple *I–V* model is presented for a MOSFET that operates in the saturation regime. The reader is referred to [5] for a complete model that is applicable to both linear and saturation regimes and the transition between the two. The drain current normalized to the width, I_D/W, of a MOSFET can be described by the product of the local charge areal density times the local carrier velocity anywhere in the channel. According to the "virtual source" MOSFET model, for a well-tempered MOSFET, the inversion charge density at the virtual source is independent of the drain voltage except for the DIBL. Also, by direct comparison to measured data, it turns out that the average carrier velocity at the virtual source is weakly dependent on either V_{GS} or V_{DS}, provided that the transistor is biased in saturation. Hence,

$$I_D/W = Q_{ix0} v_{x0} \tag{1.1}$$

where, Q_{ix0} is the inversion charge density and v_{xo} is the average carrier velocity at the virtual source. To model the virtual-source inversion charge density, the following empirical expression allows for a continuous model from weak to strong inversion:

$$Q_{ix_0} = C_{\text{inv}}\, n\phi_t\, \ln\left(1 + \exp\frac{V'_{GS} - (V_T - \alpha\phi_t F_f)}{n\phi_t}\right) \tag{1.2}$$

where C_{inv} is the effective gate-to-channel capacitance per unit area in strong inversion, ϕ_t is the thermal voltage (k_BT/q), $V'_{GS} = V_{GS} - I_D\, R_S$ is the internal gate–source voltage, i.e., corrected for the voltage drop on the source resistance R_S, n is the subthreshold coefficient, which is related to the so-called "subthreshold swing" by $S = n\,\phi_t \ln 10$, and $V_T = V_{T0} - \delta\, V'_{DS}$, where V_{T0} is the strong-inversion threshold voltage at $V_{DS} = 0$, and $V'_{DS} = V_{DS} - I_D\ (R_S + R_D)$ is the internal source–drain voltage corrected for the drop across the source and drain resistances, with δ being the DIBL coefficient with units of volt per volt. The term $\alpha\,\phi_t F_f$ allows for the requirement of different values of the threshold voltage in strong and weak inversion or in other words the so-called "constant-current" and "extrapolated" threshold voltages. A Fermi transition function F_f is used to allow for smooth transition between the two values [5].

In the strong inversion the above model can be reduced to

$$I_D/W = C_{\text{inv}}(V_{GS} - V_T)v, \tag{1.3}$$

where

$$v = \frac{v_{x0}}{1 + C_{\text{inv}} R_S W\,(1 + 2\delta)v_{x0}} \tag{1.4}$$

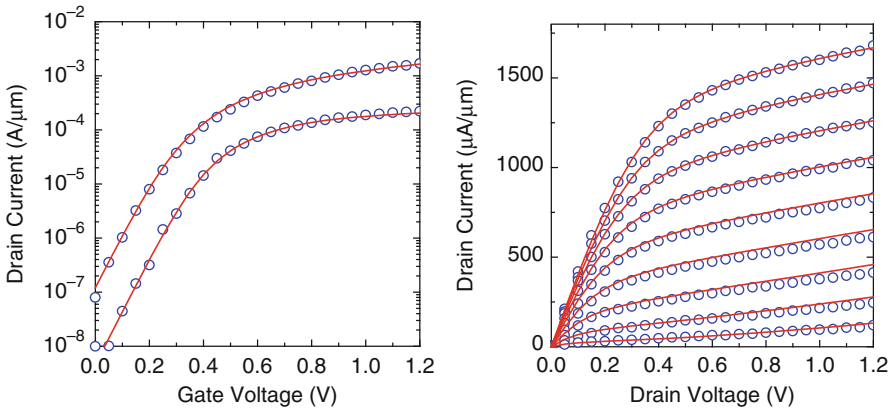

Fig. 1.4 Comparison of the analytical transistor model (*lines*) with data (*circles*) for a 65-nm node uniaxially strained NFET

is the effective velocity. Note that the notion of the "effective velocity" is just to simplify the mathematics. In the presence of S/D series resistance the internal gate-source voltage is less than the V_{GS} measured at the terminals and hence the actual inversion charge is less than $C_{\text{inv}}(V_{GS} - V_T)$. The effective velocity is defined as the apparent velocity for (1.3) to hold as if the inversion charge is given by $C_{\text{inv}}(V_{GS} - V_T)$.

1.5 MOSFET Performance Metric

Historically a *CV/I* metric has been used as a measure of the intrinsic MOSFET performance, where C is usually the inversion capacitance, i.e. $C_{\text{inv}} L_G$, sometimes corrected for some parasitic capacitances, V is the operating voltage V_{DD}, and I is the on-current, i.e., the current at $V_{GS} = V_{DS} = V_{DD}$. This simple metric worked fine for earlier technology nodes despite the fact that the switching charge $C_{\text{inv}} L_G V_{DD}$ does not include the parasitic capacitances inherent to the transistor and that the drain voltage never reaches I_{on} during switching. This is because of the fact that this metric also overestimates the inversion charge. However, as stated in Sect. 1.1, the relative importance of parasitic capacitances increases as transistor dimensions shrink.

To refine the delay metric, an "effective current", defined as the average of several points on the *I–V* loci of the transistor during the transition between logic states, was shown to better predict transistor delay. The most common definition is [6]:

$$I_{\text{eff}} = (I_D(V_{GS} = V_{DD}, V_{DS} = V_{DD}/2) + I_D(V_{GS} = V_{DD}/2, V_{DS} = V_{DD}))/2 \quad (1.5)$$

A significant observation is that usually the ratio $I_{\text{eff}}/I_{D\text{sat}}$ decreases as transistors are scaled down, mainly due to increased DIBL, which decreases the output resistance of the transistor. Future device designs should thus be aimed at increasing the effective current by controlling the short channel effects, while maintaining an acceptable on-current.

The intrinsic transistor delay is defined as $\tau = \Delta Q_G/I_{\text{eff}}$ [3], where I_{eff} is given by (1.5) and ΔQ_G is the charge difference between the two logic states, that includes both channel and fringing field charges. It follows that:

$$\tau = \frac{(1-\delta)V_{DD} - V_T + (C_f^* V_{DD}/C_{\text{inv}} L_G)}{(3-\delta)V_{DD}/4 - V_T} \frac{L_G}{v} \tag{1.6}$$

The above delay formulation should be compared to the conventional CV/I:

$$\tau = \frac{V_{DD}}{V_{DD} - V_T} \frac{L_G}{v} \tag{1.7}$$

that shows no dependence on DIBL and parasitic capacitances.

Some comments are in order here: First, the delay formulation of (1.6) uses the concept of effective current, which is only valid for $V_{DD} > 2V_T$ [6]. Also, strictly speaking, the delay formulation should use PMOS parameters in the numerator for charge estimation and NMOS parameters in the denominator for effective current calculation, and vice versa. However, in a given technology, NMOS and PMOS transistors usually have similar threshold voltages, DIBL, gate lengths, and inversion and fringing capacitances, and hence it is reasonable to use the intrinsic transistor delay given by (1.6) based on one transistor type only. The strength of (1.6) lies in the fact that it provides an analytical expression for the transistor delay based on the parameters that can be easily extracted for a given technology and have physical meaning. Therefore, it provides an easy method to explore design space, device structures, and material systems.

1.6 Historical Trend of MOSFET Performance Scaling

Figure 1.5 shows the historical trend of the intrinsic delay for some benchmark technologies, calculated using (1.6). It is interesting to note that across many technology generations with different flavors of the device architecture, the intrinsic transistor delay has scaled almost linearly in proportion to the gate length. Of course, in recent years various strain engineering methods have been incorporated to enhance carrier transport in the channel to continue the historical scaling trend. As reflected in Fig. 1.5, strain engineering is in fact essential for continued performance increase, otherwise there would be saturation in the delay versus gate length behavior.

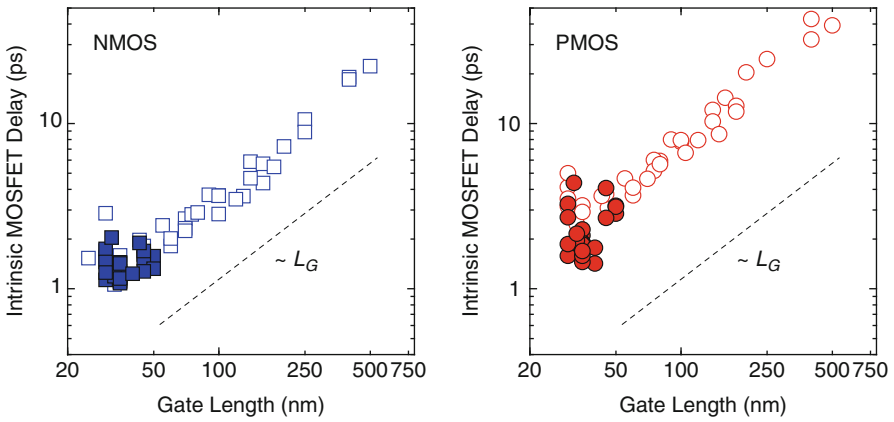

Fig. 1.5 Historical trend of the intrinsic transistor delay for some benchmark technologies. *Filled symbols* represent strain-engineered devices. Across many technology generations with different device designs, the intrinsic transistor delay has scaled almost linearly in proportion to the gate length. However, saturation of performance scaling is seen in the most recent technology nodes, especially if strain engineering is not used

1.7 Velocity Evolution in Silicon-Based MOSFETs

Even though the relative importance of the parasitic components, mainly parasitic capacitance through the increase in the $C_f^*/C_{\text{inv}}L_G$ term in (1.6), grows as the transistors are scaled, Fig. 1.5 shows that over the past two decades the intrinsic MOSFET delay has decreased in proportion to the gate length. In fact, to compensate for the increase in the first term of (1.6) and maintain commensurate scaling of the delay with gate length, the effective velocity had to increase. To analyze the evolution of velocity with dimensional scaling, it is more instructive to perform the study in terms of the virtual source velocity, v_{x0}. The virtual source velocity is in-turn related to the ballistic velocity, v_θ, through the ballistic efficiency, B,

$$v_{x0} = Bv_\theta = \frac{\lambda}{2l + \lambda} v_\theta \tag{1.8}$$

where λ is the backscattering mean free path of carriers in the vicinity of the virtual source and l is the critical length for backscattering to the source [7], which is shown through Monte Carlo simulations to be proportional to the distance over which the potential drops by k_BT/q.

Figure 1.6 shows the extracted virtual source velocity for the benchmark technologies as a function of gate length. As l decreases in proportion to the channel length, the virtual source velocity increases. However, for gate lengths below 100 nm there is a saturation in the velocity, most likely due to increased Coulomb

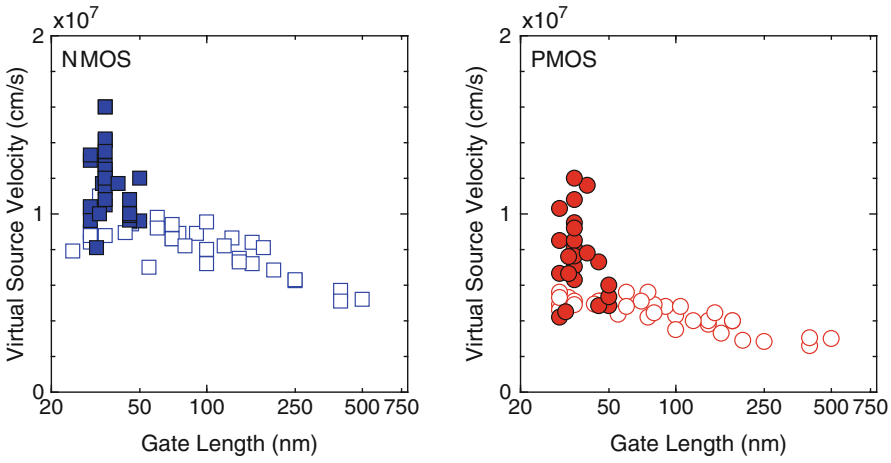

Fig. 1.6 Extracted virtual source velocity, v_{x0}, as a function of gate length for benchmark technologies. *Filled symbols* represent strain-engineered devices. Saturation of virtual source velocity is seen for unstrained devices below a gate length of about 100 nm

scattering that results from increased doping necessary to maintain electrostatic integrity. In recent years, innovations in strain-engineering have restored the velocity increase by improving mobility and ballistic velocity.

1.8 Limits of Velocity Increase in Strained Silicon

So far, electron velocity enhancement has been limited by the magnitude of uniaxial strain exerted on the transistor channel. While with most local strain-engineering methods, channel strain is limited to about 0.5%, it is possible to achieve uniaxial strain with a magnitude well beyond 1% by preferential relaxation of biaxial strain [8–10]. Early short-channel devices show promising results [10], but devices with more competitive S/D resistances are yet to be fabricated to assess whether higher strain levels translate to significant enhanced carrier velocity. Nevertheless, the fact that mobility enhancement of about 100% is maintained at low temperatures [8] suggests that the enhancement is most likely due to a decrease in effective mass and hence should yield significant velocity increase [3].

On the contrary, experimental data on short-channel PFETs demonstrate that although hole mobility as high as 4× with uniaxially strained (100) Si [11], and 8× with strained (110) wafers [12] compared to relaxed (100) Si, has been observed, the enhancement in virtual source velocity is limited to about a factor of 2, as shown in Fig. 1.7a. In fact, band structure calculations suggest that the enhancement in ballistic hole velocity saturates to about a factor of 2, as shown in Fig. 1.7b. This is due to the fact that the band structure near the top of the valence

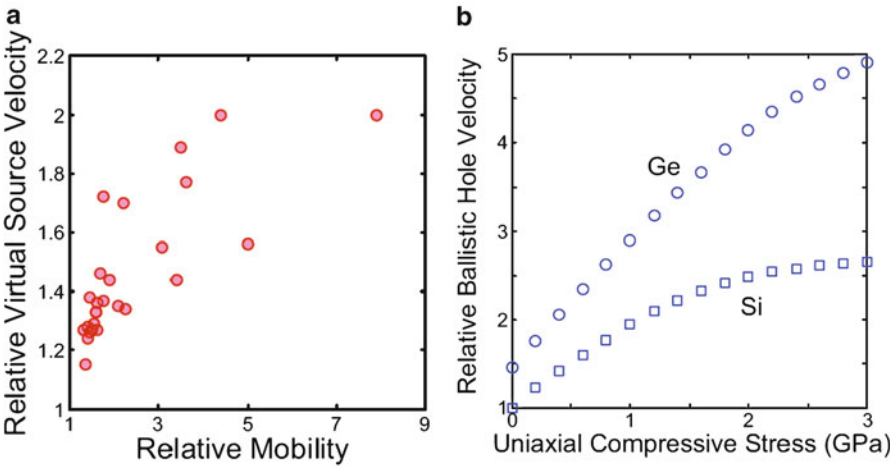

Fig. 1.7 (**a**) Relative virtual source velocity as a function of relative mobility, extracted from short channel strain-engineered PFETs. Data are referenced to control devices built on (100) Si with [110] channel direction. (**b**) Calculated ballistic hole velocity in uniaxially strained Si and Ge with (100) and [110] surface and channel orientation, respectively, and normalized to the velocity in relaxed Si. Calculations are performed with non-self-consistent $k \cdot p$ method and demonstrate that in strained Si ballistic hole velocity does not increase with compressive stress above roughly 2 GPa. Relaxed Ge does not offer significant benefits over Si, in agreement with experimental data. However, considerably higher velocities are expected with uniaxially strained Ge (Reprinted with permission from [4]. © 2005 IEEE)

band does not change further once a stress level of about 2 GPa is reached. The band structure at higher energies, which are still within reach of optical phonon energy, continues to change and thus offers higher hole mobility due to reduction in interband scattering.

1.9 Prospects of Velocity Increase in Ge and III–V Semiconductors

Figure 1.8a compares the estimated virtual-source velocity from short-channel Ge PFETs with corresponding numbers in state-of-the-art relaxed and strained Si transistors. Only a small enhancement is expected from either relaxed or biaxially strained Ge and SiGe, despite the fact that mobility enhancement in excess of a factor of 2 compared to relaxed Si has been reported in these devices. On the contrary, with uniaxially strained Ge significantly higher hole velocities are expected as shown in Fig. 1.7b. In fact, early uniaxially strained SiGe-channel MOSFETs show promising results and further advancement is anticipated in the future as embedded SiGe stressor looses its effectiveness in sub-100-nm transistor pitch and is not straightforward to be integrated with FDSOI device structures that are anticipated in 20-nm node and beyond.

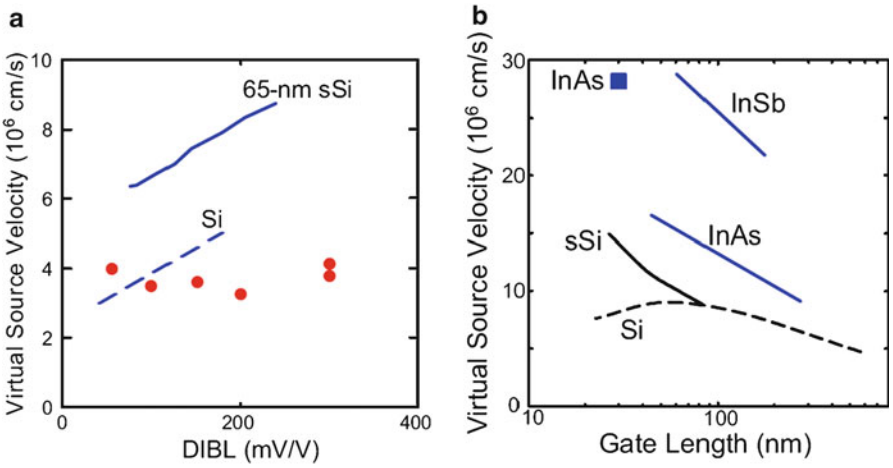

Fig. 1.8 (**a**) Hole virtual source velocity extracted from short-channel Ge PFETs from literature (*symbols*) compared with historical data for relaxed and uniaxially strained Si. Experimental data for both relaxed and biaxially strained Ge are included, demonstrating that without uniaxial strain, Ge does not offer any benefit over relaxed Si. (**b**) Comparison of virtual source velocity in deeply scaled III–V HEMTs and historical Si data (Reprinted with permission from [4]. © 2005 IEEE)

For electrons, III–V channel transistors [13–15] offer significantly higher velocity over state-of-the-art strained Si as shown in Fig. 1.8b. However, this is accompanied by small band-gap and effective mass, which lead to high band-to-band tunneling that limits the supply voltage [16]. Special device design combining the reduced-drain-electric-field benefits of HEMT architecture with reduced resistance of source/drain self-alignment will be essential to realize the intrinsic electron transport benefits. An additional challenge is that due to small quantization mass, the inversion capacitance with these materials is relatively small compared to Si MOSFETs with equivalent dielectric thickness and this can limit the anticipated drive current increase and device performance [17]. Note that since the switching charge is dominated by parasitic capacitances, the conventional *CV*/*I* metric, which suggests that reduction of the inversion capacitance is beneficial, is not a good choice for performance benchmarking particularly in this case [3]. Thus, it is anticipated that III-V channel materials can only have significant NFET performance benefit if T_{inv} (thickness of inversion layer) and R_S values are similar to competing Si devices. The same requirements also hold for PFETs with Ge or GeSi compounds, which do hold promise of higher hole velocity than Si.

1.10 Summary and Conclusions

This chapter provided an overview of the historical MOSFET performance scaling and studied basic dependencies of the performance on device parameters. Basics of MOSFET operation in deca-nanometer regime were reviewed and a simple

analytical model was presented that describes the *I–V* characteristics of the MOSFET based on a set of parameters that are physically meaningful and are usually reported in the literature. An analytical expression for intrinsic MOSFET delay was then provided and used to study the historical trends of MOSFET performance scaling. Virtual source velocity of carriers was shown to be the main driver for increased performance over the past two decades. Limits of strain-engineered velocity increase in Si-channel MOSFETs were discussed and it was concluded that it is unlikely that strained-Si provides velocity increase higher than what is achieved in today's state-of-the-art MOSFETs. Finally, prospects of velocity increase in Ge and III–V channel devices were discussed. It was shown that without uniaxial compressive strain, Ge or GeSi-channel MOSFETs do not provide any benefit over strained-Si. Higher velocities are expected in uniaxially strained Ge, but solid experimental data is still missing. In contrast, electron velocity values significantly higher than achieved in strained-Si have been already demonstrated in III-V devices. However, the performance of these devices is mainly limited because of the small quantization and density-of-states effective mass in these materials, which results in smaller C_{inv} compared to Si MOSFET.

The material presented in this chapter provides the basis for benchmarking devices with novel channel materials. In particular, (1.2)–(1.4) provide a methodology for extracting virtual source velocity from *I–V* and *C–V* measurements. As long as the measured transistors are well-behaved, i.e., have reasonable DIBL, and have a gate length less than about 100 nm, the extracted velocity can be used to predict the performance of a hypothetical device with more aggressive gate length and with aggressively scaled gate dielectric and reduced series resistance. Equation (1.6) then gives an estimate of the intrinsic delay of such a hypothetical transistor provided that realistic assumptions are made about the parasitic capacitances. This approach yields more realistic benchmarking results than using popular metrics such as long channel mobility, on-current, or delay and energy calculations that are based on the *CV/I* metric.

References

1. R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. Leo Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. 9, pp. 256–268, 1974.
2. F. Boeuf, *et al.*, "0.248 μm^2 and 0.334 μm^2 conventional bulk 6 T-SRAM bit-cells for 45 nm node low cost - general purpose applications," in *Symp. VLSI Tech.*, pp. 130–131, 2005.
3. A. Khakifirooz and D. A. Antoniadis, "MOSFET performance scaling – Part I: Historical trends," *IEEE Trans. Electron Devices*, vol. 55, no. 6, pp. 1391–1400, 2008.
4. A. Khakifirooz and D. A. Antoniadis, "MOSFET Performance scaling – Part II: Future directions," *IEEE Trans. Electron Devices*, vol. 55, no. 6, pp. 1401–1408, 2008.
5. A. Khakifirooz, O.M. Nayfeh, and D.A. Antoniadis, "A simple semiempirical short-channel MOSFET current–voltage model continuous across all regions of operation and employing only physical parameters," *IEEE Trans. Electron Devices*, vol. 56, no. 8, pp. 1674–1680, 2008.
6. M. H. Na, E. J. Nowak, W. Haensch, and J. Cai, "The effective drive current in CMOS inverters," in *IEDM Tech. Dig.*, Dec. 2002, pp. 121–124.

7. M. Lundstrom, "On the mobility versus drain current relation for a nanoscale MOSFET," *IEEE Electron Device Lett.*, vol. 22, no. 6, pp. 293–295, 2001.
8. T. Irisawa, T. Numata, T. Tezuka, N. Sugiyama, and S. Takagi, "Electron transport properties of ultrathin-body and tri-gate SOI nMOSFETs with biaxial and uniaxial strain," in *IEDM Tech. Dig.*, 2006, pp. 457–460.
9. P. Hashemi, L. Gomez, M. Canonico, and J.L. Hoyt, "Electron transport in gate-all-around uniaxial tensile strained-Si nanowire n-MOSFETs," in *IEDM Tech. Dig.*, 2008, pp. 865–868.
10. K. Maitra, A. Khakifirooz, P. Kulkarni, *et al.*, "Aggressively scaled strained-silicon-on-insulator undoped-body high-k/metal-gate nFinFETs for high-performance logic applications," *IEEE Electron Device Lett.*, vol. 32, no. 6, pp. 713–715, 2011.
11. S. Narasimha, *et al.*, "High performance 45 nm SOI technology with enhanced strain, porous low-k BEOL, and immersion lithography," in *IEDM Tech. Dig.*, 2006, p. 689.
12. B. Yang, *et al.*, "Stress dependence and poly-pitch scaling characteristics of (110) PMOS drive current," in *Symp. VLSI Tech.*, 2007, pp. 126–127.
13. D.-H. Kim and J. A. del Alamo, "Logic Performance of 40 nm InAs HEMTs," in *IEDM Tech. Dig.*, 2007, p. 629.
14. D.-H. Kim and J. del Alamo, "30 nm E-mode InAs PHEMTs for THz and future logic applications," in *IEDM Tech. Dig.*, 2008, p. 30.1.1.
15. G. G. Dewey, M. K. Hudait, K. Lee, R. Pillarisetty, W. Rachmady, M. Radosavljevic, T. Rakshit, and R. Chau, "Carrier transport in high-mobility III–V quantum-well transistors and performance impact for high-speed low-power logic applications," *IEEE Electron Device Lett.*, vol. 29, no. 10, pp. 1094–1097, 2008.
16. D. Kim, T. Krishnamohan, H.S.P. Wong, and K.C. Saraswat, "Band to band tunneling study in high mobility materials : III-V, Si, Ge and strained SiGe," in *Device Research Conf.*, 2007, p. 57.
17. K. D. Cantley, Y. Liu, H. S. Pal, T. Low, S. S. Ahmed, and M. S. Lundstrom, "Performance analysis of III-V materials in a double-gate nano-MOSFET," in *IEDM Tech. Dig.*, 2007, p. 113.

Further Reading

Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, 2nd Ed., 2009.
D.K. Schroder, *Semiconductor Material and Device Characterization*, John Wiley and Sons, 3rd Ed., 2006.
M. Lundstrom, *Fundamentals of Carrier Transport*, Cambridge University Press, 2000.
S. Datta, *Electronic Transport in Mesoscopic Systems*, Cambridge University Press, 1997.

Chapter 2
Electronic Transport in Graphene

Jun Zhu

This chapter provides an experimental overview of the electrical transport properties of graphene and graphene nanoribbons, focusing on phenomena related to electronics applications. Section 2.1 gives a brief description of the band structure. Section 2.2 discusses the effect of various scattering mechanisms in 2D sheets and nanoribbons and compares the characteristics of exfoliated and synthesized graphene. The physics of high-bias transport in graphene field effect transistors is described in Sect. 2.3. Section 2.4 gives a brief summary and outlook.

2.1 Electronic Band Structure of Graphene

2.1.1 Tight-Binding Calculations

Figure 2.1 shows the hexagonal lattice of graphene, where each unit cell contains two carbon atoms, A and B. The lattice vectors $a_1 = a/2(\sqrt{3}, 1)$ and $a_2 = a/2(\sqrt{3}, -1)$, where $a = \sqrt{3}a_{c-c} = 2.46\,\text{Å}$ is the lattice constant. In the reciprocal lattice space, the first Brillouin zone spanned by the wave vector $k = (k_x, k_y)$ is also hexagonal. Of particular importance are the two inequivalent points K and K′ at the six corners of the Brillouin zone. Their positions are given by $\pm 2\pi/\sqrt{3}a(1, 1/\sqrt{3})$, $\pm 2\pi/\sqrt{3}a(0, 2/\sqrt{3})$, and $\pm 2\pi/\sqrt{3}a(-1, 1/\sqrt{3})$. The low-energy electronic band structure of graphene can be well described by a simple tight-binding Hamiltonian

J. Zhu (✉)
Department of Physics, Penn State University, University Park, PA 16802, USA
e-mail: jzhu@phys.psu.edu

R. Murali (ed.), *Graphene Nanoelectronics: From Materials to Circuits*,
DOI 10.1007/978-1-4614-0548-1_2,

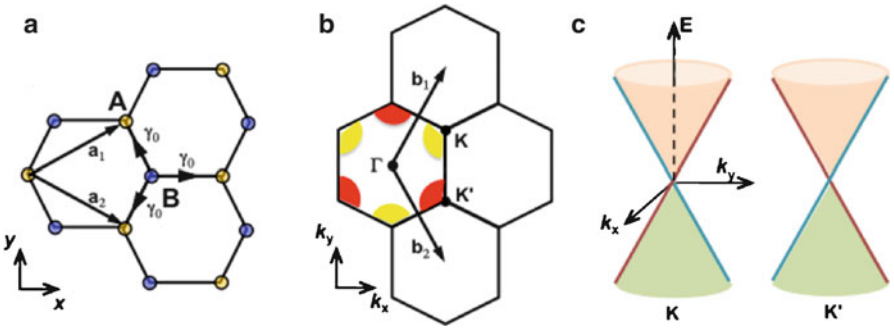

Fig. 2.1 The band structure of graphene. (**a**) The hexagonal lattice of graphene. (**b**) The reciprocal lattice in momentum space. (**c**) The Dirac cones near the K and K′ points of the Brillouin zone. Carriers in the same cone with opposite momenta have opposite pseudo-spins. Carriers in different cones with opposite momenta have the same pseudo-spin

considering only nearest-neighbor hopping and one π orbital per carbon atom. This simple model allows for an analytical solution of the energy bands [1]:

$$E^{\pm}(k_x, k_y) = \pm\gamma_0\sqrt{1 + 4\cos\left(\sqrt{3}\mathrm{k}_x\mathrm{a}\right)\ \cos\left(k_y\mathrm{a}\right) + 4\ \cos^2(k_y a/2)} \tag{2.1}$$

where $\gamma_0 \sim 2.7$ eV is the nearest-neighbor hopping integral shown in Fig. 2.1a.

In pristine undoped graphene, the conduction and valence bands touch at the K and K′ points. Expanding equ. (2.1) near K (K′) yields a linear dispersion:

$$E^{\pm}(\boldsymbol{\kappa}) = \pm\hbar v_F|\kappa| \tag{2.2}$$

where $\boldsymbol{\kappa} = \boldsymbol{k}-\boldsymbol{K}$ ($\boldsymbol{K}'$) is the wave vector measured from $\boldsymbol{K}$ ($\boldsymbol{K}'$) and v_F is the electronic group velocity given by:

$$v_F = \sqrt{3}\gamma_0 a/2\hbar \tag{2.3}$$

Equation (2.2) produces touching conic bands at the K and K′ points of the Brillouin zone as shown in Fig. 2.1c. Electrons in graphene are referred to as Dirac fermions because of this linear, photon-like dispersion and the touching points in momentum space are called Dirac points. The Fermi level E_F resides precisely at the Dirac point in undoped graphene. E_F increases (decreases) upon electron (hole) doping. Equation (2.2) is a good approximation as long as the energy does not deviate too far from E_F, or conversely that the momentum does not deviate too far from the K (K′) point. This condition is satisfied in most current graphene devices.

Angle resolved photoemission spectroscopy (ARPES) shows that electron–phonon coupling produces small deviations from the linear bands as E_F rises [2] above 200 meV. Graphene also possesses an additional quantum number called pseudo-spin, which originates from its two equivalent A and B sublattices. The pseudo-spin of an electron rotates along the Dirac cone, locked to the momentum wave vector $\boldsymbol{k}$.