Transport Enhancement Techniques for Nanoscale MOSFETs

by

Ali Khakifirooz

M.Sc., Electrical Engineering
University of Tehran, 1999

B.Sc., Electrical Engineering
University of Tehran, 1997

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2008

Author .....
Department of Electrical Engineering and Computer Science
January 25, 2008

Certified by .....
Dimitri A. Antoniadis
Ray and Maria Stata Professor
Thesis Supervisor

Accepted by .....
Terry P. Orlando
Professor of Electrical Engineering
Chairman, Department Committee on Graduate Students

Transport Enhancement Techniques for

Nanoscale MOSFETs

by

Ali Khakifirooz

Submitted to the Department of Electrical Engineering and Computer Science
on January 25, 2008, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering

Abstract

Over the past two decades, intrinsic MOSFET delay has been scaled commensurate with the scaling of the dimensions. To extend this historical trend in the future, careful analysis of what determines the transistor performance is required. In this work, a new delay metric is first introduced that better captures the interplay of the main technology parameters, and employed to study the historical trends of the performance scaling and to quantify the requirements for the continuous increase of the performance in the future. It is shown that the carrier velocity in the channel has been the main driver for the improved transistor performance with scaling. A roadmapping exercise is presented and it is shown that new channel materials are needed to lever carrier velocity beyond what is achieved with uniaxially strained silicon, along with dramatic reduction in the device parasitics. Such innovations are needed as early as the 32-nm node to avoid the otherwise counter-scaling of the performance.

The prospects and limitations of various approaches that are being pursued to increase the carrier velocity and thereby the transistor performance are then explored. After introducing the basics of the transport in nanoscale MOSFETs, the impact of channel material and strain configuration on electron and hole transport are examined. Uniaixal tensile strain in silicon is shown to be very promising to enhance electron transport as long as higher strain levels can be exerted on the device. Calculations and analysis in this work demonstrate that in uniaxially strained silicon, virtual source velocity depends more strongly on the mobility than previously believed and the modulation of the effective mass under uniaxial strain is responsible for this strong dependence. While III-V semiconductors are seriously limited by their small quantization effective mass, which limits the available inversion charge at a given voltage overdrive, germanium is attractive as it has enhanced transport properties for both electrons and holes. However, to avoid mobility degradation due to carrier confinement as well as $L - \Delta$ interband scattering, and to achieve higher ballistic velocity, (111) wafer orientation should be used for Ge NFETs.

Further analysis in this work demonstrate that with uniaixally strained Si, hole

ballistic velocity enhancement is limited to about 2×, despite the fact that mobility enhancement of about 4× has been demonstrated. Hence, further increase of the strain level does not seem to provide major increase in the device performance. It is also shown that relaxed germanium only marginally improves hole velocity despite the fact that mobility is significantly higher than silicon. Biaxial compressive strain in Ge, although relatively simple to apply, offers only 2× velocity enhancement over relaxed silicon. Only with uniaxial compressive strain, is germanium able to provide significantly higher velocities compared to state-of-the-art silicon MOSFETs.

Most recently, germanium has manifested itself as an alternative channel material because of its superior electron and hole mobility compared to silicon. Functional MOS transistors with relatively good electrical characteristics have been demonstrated by several groups on bulk and strained Ge. However, carrier mobility in these devices is still far behind what is theoretically expected from germanium. Very high density of the interface states, especially close to the conduction band is believed to be responsible for poor electrical characteristics of Ge MOSFETs. Nevertheless, a through investigation of the transport in Ge-channel MOSFETs and the correlation between the mobility and trap density has not been undertaken in the past. Pulsed $I-V$ and $Q-V$ measurement are performed to characterize near intrinsic transport properties in Ge-channel MOSFETs. Pulsed measurements show that the actual carrier mobility is at least twice what is inferred from DC measurements for Ge NFETs. With phosphorus implantation at the Ge-dielectric interface the difference between DC and pulsed measurements is reduced to about 20%, despite the fact that effects of charge trapping are still visible in these devices.

To better understand the dependence of carrier transport on charge trapping, a method to directly measure the inversion charge density by integrating the S/D current is proposed. The density of trapped charges is measured as the difference between the inversion charge density at the beginning and end of pulses applied to the gate. Analysis of temporal variation of trapped charge density reveals that two regimes of fast and slow charge trapping are present. Both mechanisms show a logarithmic dependence on the pulse width, as observed in earlier literature charge-pumping studies of Si MOSFETs with high-κ dielectrics. The correlation between mobility and density of trapped charges is studied and it is shown that the mobility depends only on the density of fast traps. To our knowledge, this is the first investigation in which the impact of the fast and slow traps on the mobility has been separated.

Extrapolation of the mobility-trap relationship to lower densities of trapped charges gives an upper limit on the available mobility with the present gate stack if the density of the fast traps is reduced further. However, this analysis demonstrates that the expected mobility is still far below what is obtained in Si MOSFETs. Further investigations are needed to analyze other mechanisms that might be responsible for poor electron mobility in Ge MOSFETs and thereby optimize the gate stack by suppressing these mechanisms.

Thesis Supervisor: Dimitri A. Antoniadis
Title: Ray and Maria Stata Professor

Acknowledgments

I would like to thank, first and foremost, my adviser, Professor Dimitri A. Antoniadis for his guidance, encouragement, and continuous support throughout my research. I owe a lot of my professional development to Dimitri for his critical thinking and insights as well as giving me the opportunity to explore so many, seemingly unrelated, projects.

I would also thank the rest of my thesis committee, Professor Jesus del Alamo and Professor Judy L. Hoyt. They offered a lot of great insights that helped me better organize this thesis and strengthen the discussions.

I would like to acknowledge the staff of the Microsystems Technology Laboratories (MTL), for their restless efforts to keep the lab running and for helping me with my device fabrication. I would like to specifically thank the staff members I personally worked with: Bernard Alamariu, who helped a lot with making germanium processing possible and developing several recipes, Dan Adams, Brian McKenna, Bob Bicchieri, Vicky Diadiuk, Eric Lim, Gary Riggott, Dave Terry, Paul Tierney, and Paudely Zamora.

I learned a great deal from the past and present members of the DAA group: Dr. Anthony Lochtefeld, Dr. Keith Jackson, Dr. Ihsan Djomehri, Professor Ganesh Samudra, Dr. Hitoshi Wakabayashi, Dr. Hasan Nayfeh, Dr. Isaac Lauer, Dr. Jongwan Jung, Dr. Andy Ritenour, Dr. Larry Lee, Dr. Scott Yu, John Hennessy, Osama Nayfeh, Jae-Kyu Lee, and Dr. Ryan Lee. Isaac was my office-mate for many years and helped me get started in the lab. Andy and I shared many failures of the early days of processing germanium wafers. Some of the measurements reported in this thesis were also carried out on devices fabricated by him and John. Osama was a great help with simulations and especially in setting up the Monte Carlo simulations of the inverse-modeled MOSFETs.

My sincere gratitude goes to Professor Mark Lundstrom, Professor Gerhald Klimeck, Dr. Anisur Rahman, and Neophytou Neophytos of Purdue University for their collaboration in the theoretical studies and providing me with the latest versions of their

simulation codes.

I had great days with the people on the sixth floor of MTL: Jim Fiorenza, John Kymissis, Andy Fan, Niamh, Cait, Joyce, Tonya, Ingvar, Nisha, Tan, Maggie, Anita, Muyiwa, Leo, Pouya, Dae-Hyun, Luis, Dennis, MK, and Nicole; I would like to thank them and the rest of MTL for their social and technical support.

I would also like to express my appreciation to Dimitri's assistants in the past and present: Bogusia Gatarz, Rose Previte, and Michele Hudak. They were great help with administrative tasks.

My research at MIT would not be possible without the funding from Semiconductor Research Corporation (SRC) and FCRP Materials, Structures, and Devices Focus Center. The annual reviews, weekly teleseminars, and other forums provided a unique opportunity to exchange ideas and receive timely comments and advice.

I owe my early acquaintance with the device physics and fabrication technology to Professor Ali Afzali-Kusha and Professor Shamsoddin Mohajerzadeh of the University of Tehran. I was also very fortunate that Dr. Ghavam Shahidi introduced me to Dimitri and always kept an eye on my progress. I would like to acknowledge their support and mentorship.

My parents, receive my deepest gratitude and love for their dedication and many years of support. My siblings, Mehdi, Marzieh, and Mohammad were a continuous source of love and I wish them success in their endeavor and happiness throughout their life. I am grateful also to my parents-in-law for being always understanding and helpful. I am also indebted to my in-laws, Mahshid, Mahdokht, and Mahdi for their encouragement and support.

Last, but not least, I would like to thank my wife Mahnaz for her extraordinary patience, understanding, and love during the past few years. Her support and encouragement was in the end what made this dissertation possible. Our little star, Setareh, typed a bit here and there in my thesis and correcting her edits often stimulated me to do more revisions. I am looking forward to returning her favor!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The main driver of the IC industry over the past four decades or so has been Moore's law, which states that the number of the transistors on a chip doubles every two years [1]. The key enabler for this exponential growth is the shrinking of the MOSFET pitch by a factor of 0.7 per technology generation. Other transistor dimensions have been scaled with almost the same pace, according to Dennard's scaling theory [2]. Up to now, MOSFET scaling has brought about a commensurate increase in the transistor speed as reflected in Figure 1-1. However, parasitic components associated with a MOSFET do not follow the same scaling trend; Their relative importance increases as the transistors are shrunk and they can ultimately hinder the performance improvement offered by the geometric scaling.

Figure 1-2 illustrates the main parasitic components associated with a state-of-the-art MOSFET. 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 as will be discussed in Chapter 2. In fact, as will be shown in the next chapter, due to the proximity of the source/drain contact studs to the gate electrode, future technology nodes will suffer from a larger amount of parasitic capacitance. This additional component is denoted by C_{pp} in Figure 1-2. 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 extensions, is about 80 $\Omega.\mu$m per side for modern NFETs

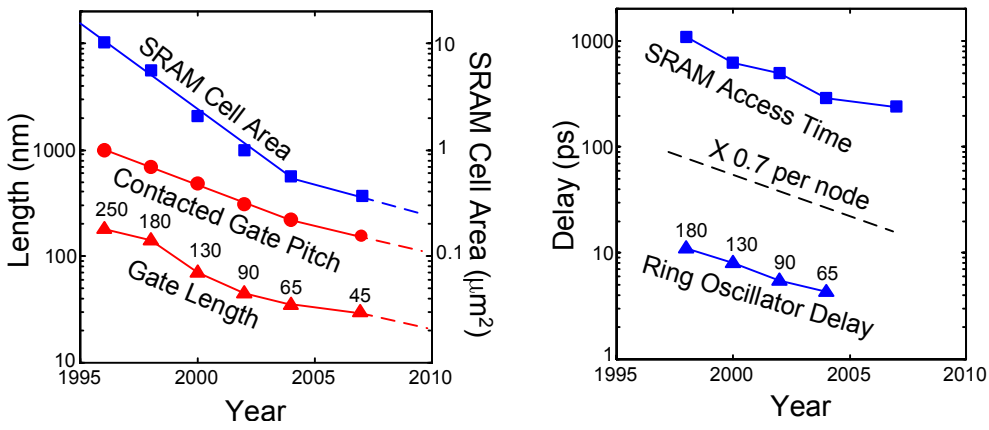


Figure 1-1: (a) Scaling trend of the MOSFET geometry over the past decade and (b) the associated decrease in the delay. Transistor pitch is scaled by a factor of 0.7 per technology node to accommodate for the doubling of the transistor count. Up to the 65-nm node, circuit delay has almost followed the same pace of scaling by a factor of 0.7 per node. (Data from Intel [4–13].)

and does not scale very well. Furthermore, in future technology nodes, the series resistance due to the finite conductivity of the metal contacts might be significant and remedies are sought to reduce this additional component [3].

An equally important issue for state-of-the-art CMOS technologies, is the scaling of the supply voltage to keep the power consumption under control as well as to maintain the device reliability [15]. The threshold voltage followed a similar scaling trend in the earlier technologies. However, for gate lengths below 100 nm, the threshold voltage scaling has to be slowed down to control the exponentially growing standby power [16]. This means that less gate voltage overdrive, $V_{DD} - V_T$, is available as the devices are scaled. In conjunction with the increased importance of the parasitic components, this decelerates the performance scaling as will be discussed in Chapter 2.

A possible solution to compensate for the performance drop imposed by the loss of the gate overdrive and the relative increase of the parasitic components, is to improve transport properties of the channel by employing new materials. Strained silicon grown pseudomorphically on SiGe virtual substrates is a classical candidate [17]. Strained silicon directly on insulator (SSDOI) has been proposed to enjoy from the

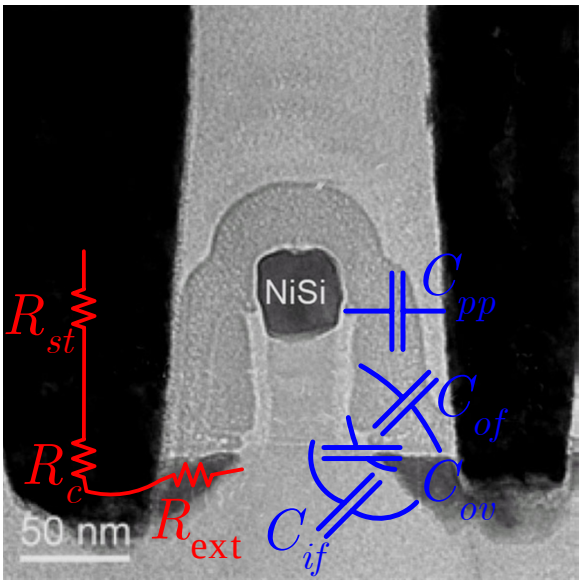


Figure 1-2: The main parasitic components associated with a modern MOSFET: The effective fringing capacitance, C_f^*, which consists of the inner fringe, C_{if}, outer fringe, C_{of}, and overlap capacitance, C_{ov} is approximately 0.5 fF/μm and does not scale with the gate length. In fact, due to the proximity of the gate electrode and S/D studs, another term, C_{pp}, is added as the devices are 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 the 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 picture is courtesy of STMicroelectronics [14] and shows a 45-nm node MOSFET*.)

*In fact the contact pitch of the device shown here is not as aggressive as projected by Figure 1-1 for a 45-nm technology. The TEM picture here serves for the illustration purpose only.

short-channel effect immunity offered by ultrathin body SOI [18, 19]. However, once the body thickness is thinned below 3-4 nm, as required for the devices with a gate length of 10 nm, it is doubtful that biaxially-strained silicon offers any benefit over relaxed silicon [20], since in ultrathin relaxed Si channels also practically all of the electrons are in the Δ_2 band. Furthermore, as will be discussed in Chapter 3, the electron mobility enhancement observed in biaxially strained Si does not translate to the same amount of increase in the drive current and hole mobility enhancement is only seen at low gate voltages [21].

Strained-silicon/SiGe heterostructures either on bulk or on insulator (HOI) [22] are alternative options to deliver superior hole transport properties. However, scalability of such structures for deeply-scaled MOSFETs has not been demonstrated yet.

A variety of methods to induce uniaxial strain in the channel via various process-controlled mechanisms have been already implemented in manufacturing and demonstrated to significantly improve the transistor performance [9–11, 23, 24]. Along with different channel [25, 26] and wafer orientations [27–29], this offers a viable method to improve the device performance. However, despite the early success of such approaches, scalability of these methods to future technology nodes and higher device performance is subject of uncertainty and of ongoing research [30].

More exotic channel materials, such as III-V semiconductors, are being explored by several researchers in order to benefit from their superior electron mobility despite many process challenges [31, 32]. In fact, superior transport properties in III-V materials have always fascinated the semiconductor community to pursue possibilities to fabricate MOSFETs with a III-V channel. As the silicon MOSFET scaling approaches its limit, there is an increasing desire to explore new channel materials in order to extend MOSFET performance scaling in the next decade.

To assess the performance enhancement offered by each of the above options and to identify the best combination for a given technology node, careful analysis of the device structure and material system is required. This thesis aims to explore some of the potential candidates and identify their prospects and limitations. This thesis will focus on the impact of the innovative material systems on the device

performance under the assumption that process challenges can be eventually solved. In doing so, first an analytical performance metric is developed to enable realistic benchmarking of the MOS transistors. This model is used to depict the historical trend of the MOSFET performance scaling, identify the required performance in the future nodes, and analyze the trade-offs in the device design. These observations are used to compare material and device structures proposed to improve carrier transport and thereby the transistor performance.

Most recently, germanium has manifested itself as an alternative channel material because of its superior electron and hole mobility compared to silicon. Functional MOS transistors with relatively good electrical characteristics have been demonstrated by several groups on bulk [33–37] and strained Ge [38, 39]. However, the performance of these devices is still far behind what is theoretically expected from germanium. A variety of theoretical and experimental methods are employed to gain insight into mechanisms that so far have prevented germanium MOSFETs from delivering their ideal performance. These insights are used as a guideline to solve some of the fabrication challenges of Ge MOSFETs.

1.1 Overview of This Work

Chapter 2 provides a theoretical analysis of what determines MOSFET performance. An analytical expression is given for the intrinsic MOSFET delay and is used to study the historical trend of transistor performance scaling and to quantify the requirements for continuous increase of performance in future technology nodes. It is shown that carrier velocity in the channel should be increased beyond what is achievable with uniaxially strained silicon in order for commensurate performance scaling to continue its historical trend.

Chapter 3 overviews different methods to increase the carrier velocity by introducing new channel materials, applying mechanical strain to the channel, or a combination of both. Prospects and limitations of different methods are studied theoretically and by analyzing published data.

Chapter 4 presents some of the electrical characterization methods employed in this work to decouple the intrinsic transport properties of the Ge MOSFETs from non-ideal phenomena caused by imperfect gate dielectric. These methods are applicable to other alternate channel materials, such as III-V semiconductors, and provide valuable insights into what is limiting the transport properties in the channel.

Chapter 5 concludes the thesis by examining the implications of the material presented in this work for CMOS manufacturing. Special attention is paid to the limits of increased transistor performance in the future nodes that can be expected from channel strain engineering. Prospects of the germanium-channel MOSFETs in deeply scaled MOSFETs are discussed. Hopefully, this will answer the question "Is it worth it to work on Ge MOSFETs? and what is gained?"

Chapter 2

MOSFET Performance Scaling

This chapter provides a theoretical analysis of what determines MOSFET performance. An analytical expression is given for the intrinsic MOSFET delay and is used to study the historical trend of transistor performance scaling and to quantify the requirements for the continuous increase of the transistor performance in the future technology generations.

Realistic benchmarking of the transistor performance is essential to quantify technology requirements in order to continue its historical scaling trend. Unlike current and past technologies where direct measurement of different figures of merit (FOMs) is possible or they are readily obtained from circuit simulations with well-calibrated models, assessing the circuit-level performance measures for future technologies is not straightforward. Circuit simulation with predictive models [40, 41] or models calibrated with TCAD and device simulations is an option. However, it is much easier to gain physical insights if analytical expressions for the desired FOMs are available.

Historically, the transistor delay has been simply approximated by $CV_{DD}/I_{D\mathrm{sat}}$, where V_{DD} is the supply voltage and $I_{D\mathrm{sat}}$ is the drain current at $V_{GS} = V_{DS} = V_{DD}$. C represents the total load capacitance, which is usually taken equal to the intrinsic gate capacitance in inversion, C_{inv}, to obtain the intrinsic MOSFET delay for circuits dominated by transistor loads. Despite the fact that the switching charge, $C_{\mathrm{inv}}V_{DD}$, does not include gate and other parasitic capacitances inherent to the transistor, and that the drain current never reaches $I_{D\mathrm{sat}}$ during switching in a CMOS configuration

[42,43], this metric also overestimates the inversion charge, and hence it has provided acceptable results for earlier technology nodes. However, as the transistors are further shrunk, the relative importance of parasitic capacitances grows, making it unrealistic to ignore them.

Recently, it has been shown that the CV/I metric better follows the experimental inverter delay if the on-current in the denominator is replaced by an effective current, I_{eff}, representing the average switching current [42–44]. A significant observation is that usually the ratio $I_{\text{eff}}/I_{D\text{sat}}$ decreases as the transistors are scaled down, mainly due to increased drain-induced barrier lowering (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 [44].

This chapter presents an analytical expression for the intrinsic MOSFET delay, which is based on the physical models for the effective current and calculates the total gate switching charge more accurately. The proposed model is applied to published device data and the historical trend of MOSFET performance scaling is examined. It is shown that increase of the carrier velocity in the channel has been the main driver for the improved transistor performance with scaling. These observations are used to explore the tradeoffs between key device parameters in order for the commensurate scaling of the device performance with its geometrical scaling to continue the historical trend.

2.1 Simple MOSFET Analytical Model

This section provides a simple analytical model for MOSFET current, which will be used to analyze the experimental results in the next sections as well as to derive an expression for the intrinsic MOSFET delay.

The width-normalized transistor current in saturation can be expressed as:

$$I_D/W = C'_{\text{inv}}(V_{GS} - V_T)v, \tag{2.1}$$

where V_T is the saturation threshold voltage obtained by linear extrapolation of the $I_D - V_{GS}$ curve as shown in Figure 2-1, C'_{inv} is the gate capacitance per unit area and in strong inversion, and v denotes the "effective" carrier velocity[1]. The effective velocity, v, is related to the average velocity of carriers at the barrier near the source, called virtual source velocity, v_{x0}, if corrections are made for the voltage drop across the source series resistance, R_S [2] [45]:

$$v = \frac{v_{x0}}{1 + C'_{\text{inv}} R_S W (1 + 2\delta) v_{x0}}, \tag{2.2}$$

where δ is the DIBL coefficient in V/V, i.e. $\delta = \partial |V_T| / \partial V_{DS}$. The DIBL dependence denotes the fact that due to the voltage drop across source and drain series resistances, the threshold voltage increases. Eq. (2.2) assumes that $R_S = R_D$.

Eq. (2.1) is justified by the observation that in state-of-the-art MOSFETs, the saturation drain current is almost a linear function of the gate voltage, corresponding to an almost constant transconductance, as shown in Figure 2-1. In this analysis we assume that the effective velocity is independent of the gate and drain voltages [3] and that (in the absence of C-V measurements for short channel devices reported in literature) the inversion charge is simply given by $Q'_{\text{inv}} = C'_{\text{inv}}(V_{GS} - V_T)$.

Note that even elaborate estimation of the inversion charge contains some degree of uncertainty. Integrating the C-V curves measured on long channel devices and applying proper shifts due to threshold voltage roll-off, DIBL, and voltage drop across source series resistance [47] is an example. This method neglects the gate length de-

[1] We used the notion of the "effective velocity" 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 $C'_{\text{inv}}(V_{GS} - V_T)$. We define the effective velocity as the apparent velocity for (2.1) to hold as if the inversion charge is given by $C'_{\text{inv}}(V_{GS} - V_T)$ [45].

[2] In cases where S/D series resistance is not given in the literature, an upper limit can be inferred from the output characteristics of the transistor by plotting the tangential line to the $I_D - V_{DS}$ curve at high V_{GS} and low V_{DS}. Note that for state-of-the-art MOSFETs the second term in the denominator of (2.2) is about 0.2. So, even significant uncertainty in determining R_S does not change the velocity estimations considerably.

[3] According to Lundstrom's theory of MOSFET operation discussed in the next section, carrier velocity in independent of the drain voltage as long as the drain voltage is higher than a few $k_B T/q$. In the degenerate limit the ballistic velocity increases with the square root of the inversion charge density or equivalently the effective electric field, E_{eff}, while mobility drops as E_{eff} increases. Thus the virtual source velocity and in turn the effective velocity are almost constant. This has been also observed in our Monte Carlo simulations [46].

pendence of the poly depletion [48] and the spread of the C-V curve in weak inversion due to short channel effects. Direct measurement of the split C-V curve in short channel devices with proper test structures is an alternative [49]. However, according to the scattering theory of MOSFETs [50] the relevant charge for current calculation in saturation is the inversion charge at the potential barrier near the source where carriers are injected into the channel (virtual source). Hence, even if experimental C-V data on short-channel devices are available, it is not without uncertainty to calculate the actual inversion charge at the virtual source[4].

Monte Carlo simulations demonstrate that while the effective velocity increases initially as the gate voltage is increased, it is almost constant at gate voltages close to V_{DD} [46]. The above assumptions, namely inversion capacitance and effective velocity independent of the gate voltage, allow us to use the threshold voltage obtained from I-V characteristics, for charge estimation as well. It is worth noting that the threshold voltage defined here is usually 200 mV higher than what is commonly reported in the literature, which is defined at a given current in subthreshold.

The width-normalized transistor off-current is then given by:

$$I_{\text{off}}/W = I_{\text{ref}} 10^{-V_T/S^*}, \tag{2.3}$$

where S^* is the effective subthreshold swing in V/decade and I_{ref} is the numerical value of the current per unit width at the threshold voltage, $V_{GS} = V_T$, which can be found empirically. It can be observed that over different technology generations $I_{\text{ref}} = Q'_0 v_{x0}$, where $Q'_0 \approx 8 \times 10^{-8}$ C/cm^2. This agrees well with our Schrödinger-Poisson simulations that at the onset of threshold, defined by extrapolating the $Q_{\text{inv}} - V_{GS}$ characteristic, the inversion charge is approximately 10^{-7} C/cm^2. For state-of-the-art MOSFETs this definition corresponds to a current density of roughly 80 and

[4]Determination of the inversion charge at the virtual source from calibrated device simulations also involves some uncertainty. To obtain the exact shape of the potential profile along the channel and hence to determine the position of the virtual source requires self-consistent simulations with calibrated transport parameters. Also, often the location where the potential peak is located lies on a region where a steep gradient of charge from S/D extensions is present. Furthermore, the lateral coordinate of the peak depends on the distance from the oxide interface. Hence, integrating the inversion charge at the potential peak or averaging the simulated velocity is not trivial.

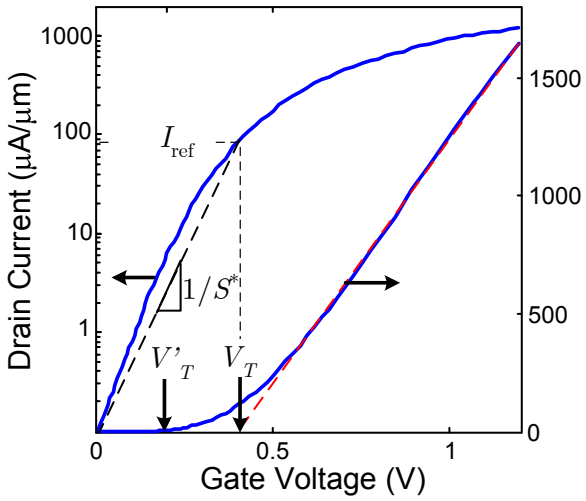


Figure 2-1: Extraction of the threshold voltage from experimental data for a 35-nm MOSFET [12] by linear extrapolation of the $I_D - V_{GS}$ at high V_{DS}. Since the velocity is expected to be lower at lower gate voltages [46], the method tends to overestimate the saturation threshold voltage for charge estimation by about 50 mV. Nonetheless, the threshold voltage defined here (V_T) is usually 200 mV higher than what is commonly reported in the literature, defined at a given current in subthreshold, (V'_T). The current at the threshold voltage is given by $I_{\rm ref} = Q'_0 v_{x0}$, where Q'_0 is empirically found to be 8×10^{-8} C/cm^2. An effective subthreshold swing, S^*, can be defined so that $I_{\rm off}/W = I_{\rm ref} 10^{-V_T/S^*}$.

40 μA/μm for NFETs and PFETs, respectively. Traditionally the threshold voltage is defined at a current numerically equal to $5 \times 10^{-7} W/L$. For a transistor with a gate length of 35 nm, this corresponds to a current density of about 14 μA/μm [5].

[5] This is in fact the definition used in ITRS calculations done in MASTAR [51]. Strictly speaking, MASTAR defines two threshold voltages, one for off-current calculations with the above definition, and one for on-current calculations, 30-50 mV above the off-state threshold voltage. MASTAR also has the option to calculate a doping-dependent threshold voltage at a current given by $5 \times 10^{-7} W/L \; 8 \times 10^8 N_A^{-0.4865}$, which only slightly differs from the traditional value. Nonetheless, ITRS does not distinguish between the two values for the on-state and off-state threshold voltage. This is one of the major shortcomings of ITRS projections and has significant consequences when quantifying technology requirements.

2.2 Performance Metric

We define the intrinsic transistor delay as $\tau = \Delta Q_G / I_{\text{eff}}$ [45], where I_{eff} is the effective drain current [42] and ΔQ_G is the charge difference between the two logic states[6], that includes both channel and fringing field charges. As detailed in Appendix A, it follows that:

$$\tau = \frac{(1-\delta)V_{DD} - V_T + (C_f^* V_{DD}/C'_{\text{inv}} L_G)}{(3-\delta)V_{DD}/4 - V_T} \frac{L_G}{v}, \tag{2.4}$$

where C_f^* represents the equivalent gate fringing capacitance, with Miller effect taken into account. The minimum possible value for C_f^* occurs for channel isoplanar with source and drain (S/D), i.e. no raised S/D or contact vias in the vicinity of the gate, and it is roughly 0.5 fF/μm for well-optimized devices and nearly independent of the technology node [52]. The above delay formulation can be compared to the conventional CV/I:

$$\frac{C'_{\text{inv}} V_{DD}}{I_{D\text{sat}}} = \frac{V_{DD}}{V_{DD} - V_T} \frac{L_G}{v}, \tag{2.5}$$

that shows no dependence on DIBL and parasitic capacitances.

Some comments are in order here: first, the delay formulation of (2.4) uses the concept of the effective current, which is only valid for $V_{DD} > 2V_T$. 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 (2.4) based on one transistor type only.

As a sanity check for the delay expression in (2.4), Figure 2-2 compares the intrinsic delay of NMOS and PMOS transistors calculated using (2.4) and the experimental ring oscillator delay for a 90-nm technology [53] as a function of the supply voltage. The inverter delay can be modeled well as the average intrinsic delay of the NMOS and

[6] One might argue that the switching charge that matters when calculating the logic propagation delay is defined between $V_{DD}/2$ points. However, we use the rail-to-rail charge difference to be consistent with the definition of the effective current.

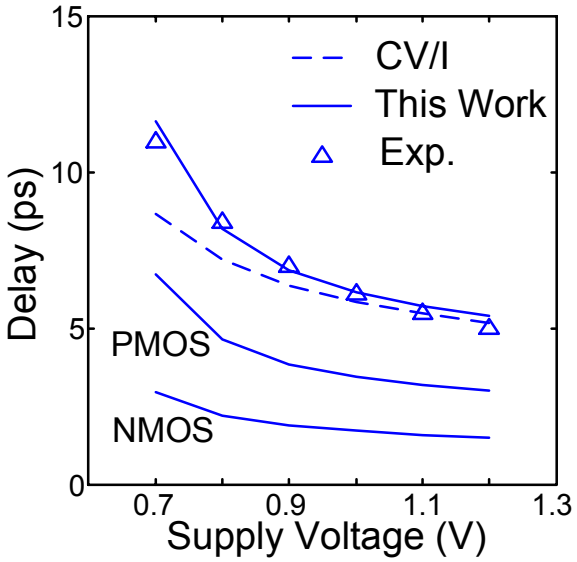


Figure 2-2: Comparison of the calculated intrinsic delay and experimental data for a 90-nm technology [53] as a function of the supply voltage. The inverter delay can be modeled well as the average intrinsic delay of the NMOS and PMOS transistors calculated using (2.4) and multiplied by a empirical scaling factor (2.2 here) to account for their unequal device widths and other parasitic capacitances, like those associated with junctions and interconnects, not included in our intrinsic delay. Conventional CV/I metric fails to model the inverter delay accurately at lower supply voltages. Note that a different scaling factor (4.5 in this case) is needed for the CV/I metric to provide values close to the actual inverter delays.

PMOS transistors calculated using (2.4) and multiplied by a scaling factor of about 2.2 to account for larger PMOS gate width. In contrast, CV/I metric fails to provide good approximations at low supply voltages even though a larger scaling factor of about 4.5 is used. Further analysis is shown in Figure 2-3, that illustrates a comparison of the measured inverter delay with the intrinsic delay calculated using (2.4) across several technology generations with various dimensions and supply voltages [5–8, 12, 53–69]. Again the proposed metric follows the experimental data very well with a constant scaling factor of 2.2, whereas the conventional CV/I metric exhibits a super-linear relationship with measured ring oscillator delay.

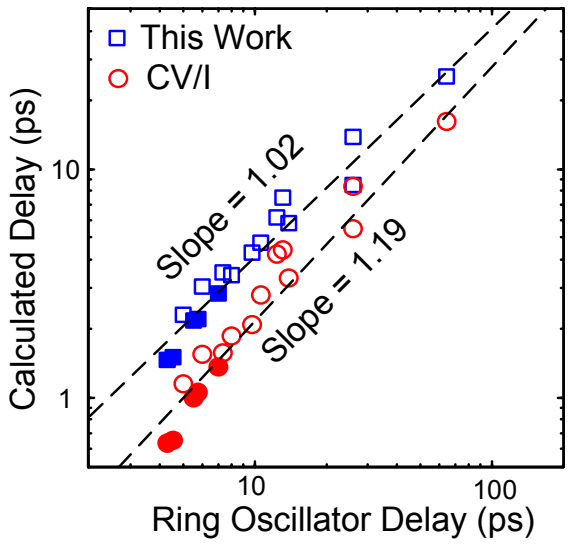


Figure 2-3: Comparison of the calculated intrinsic delay and experimental data across several technology generations [5–8, 12, 53–69]. The proposed metric follows the experimental data very well, whereas the conventional CV/I metric exhibits a super-linear relationship with measured ring oscillator delay. Filled symbols denote strain-engineered devices.

2.3 Historical Trend of MOSFET Performance Scaling

Figure 2-4 shows the historical trend of the intrinsic delay for some benchmark technologies [4–12, 14, 23, 24, 53–89], calculated using (2.4). 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 the recent years various strain engineering methods have been incorporated to enhance carrier transport in the channel in order to continue the historical scaling trend. As reflected in Figure 2-4 strain engineering is in fact essential for continued performance increase, otherwise there would be saturation in the delay versus gate length behavior.

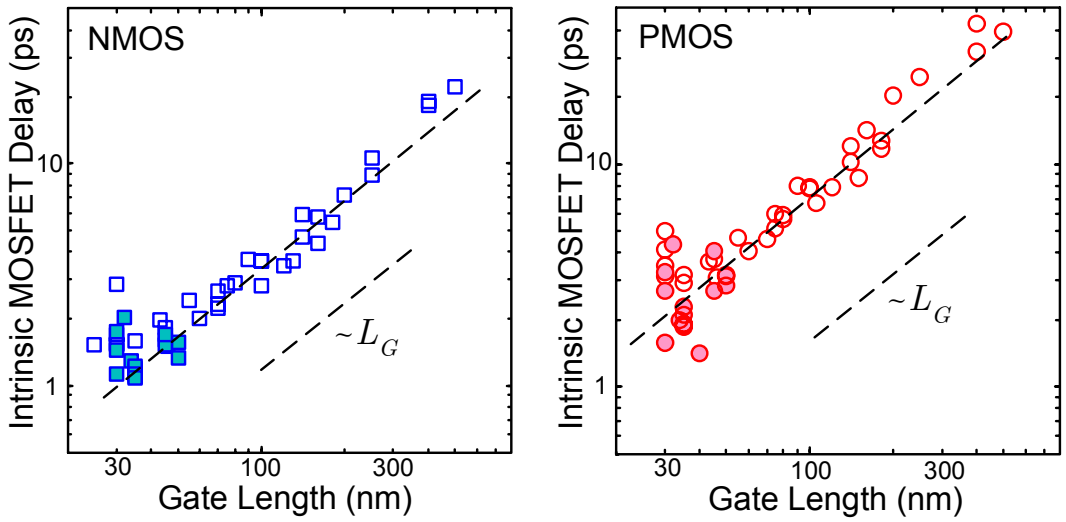


Figure 2-4: Historical trend of the intrinsic transistor delay for benchmark technologies [4–12,14,23,24,53–89]. Filled symbols represent strain-engineered devices. Across many technology generations with different flavors of the device architecture, the intrinsic transistor delay has scaled almost linearly in proportion to the gate length.

2.4 Velocity Evolution

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 (2.4), grows as the transistors are scaled, Figure 2-4 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 (2.4) and maintain commensurate scaling of the delay with gate length, the effective velocity had to increase. In order to analyze the evolution of the 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,

$$v_{x0} = Bv_\theta = \frac{\lambda}{2l+\lambda}v_\theta, \tag{2.6}$$

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 [50], which is shown through Monte Carlo simulations to be proportional to the distance over which the

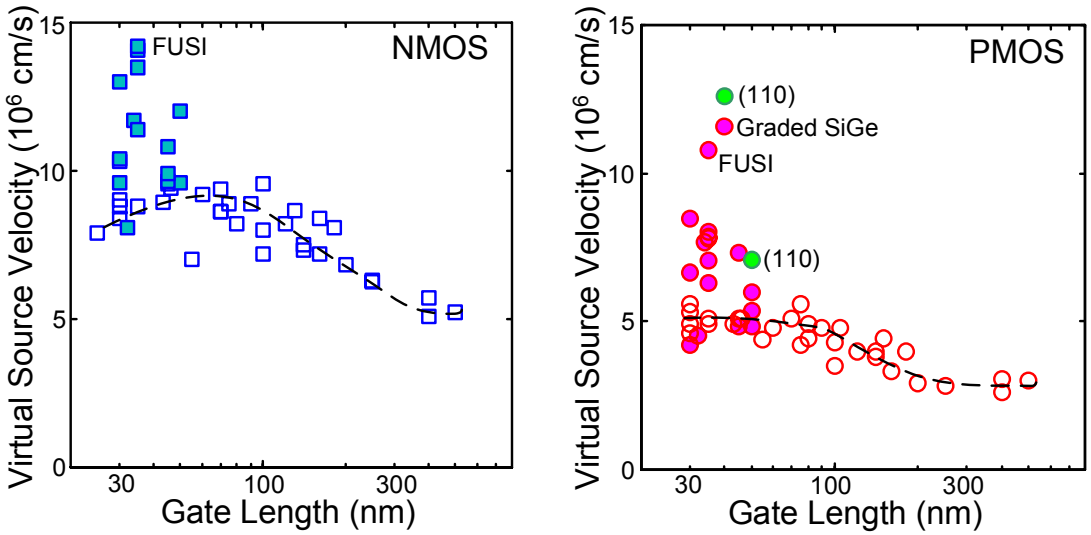


Figure 2-5: The extracted virtual source velocity, v_{x0}, as a function of the gate length for benchmark technologies [4–12, 14, 23, 24, 53–89, 92–94] . Filled symbols represent strain-engineered devices.

potential drops by k_BT/q [90, 91].

Figure 2-5 shows the extracted virtual source velocity for the benchmark technologies [4–12, 14, 23, 24, 53–89, 92–94] as a function of the gate length[7]. As l decreases in proportion to the channel length, the virtual source velocity increases. However, for gate lengths below 130 nm there is saturation in the velocity for relaxed-Si technologies, most likely due to increased Coulomb 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.

An interesting observation from Figure 2-5 is that with enhanced strain-engineering [92], through the use of new wafer orientations, or a combination of both [94], PMOS transistors are approaching NMOS devices in terms of the virtual source velocity. Even with high-κ gate dielectrics [93], more than 40 % hole velocity enhancement compared to relaxed silicon is achieved by using uniaxially strained (110) Si. As will be discussed in the next section, although hole mobility continues to increase

[7]For a fair comparison the extracted velocity should be plotted vs. effective channel length, often not reported in the literature and not easy to determine. This figure is however meant to provide the overall trend of velocity evolution and not a comparison of different technologies.

with increasing the uniaxial compressive strain on (100) wafers, it appears that velocity enhancement slows down once strain-induced modulation of the effective mass saturates. This signifies the importance of strained (110) and (111) surface orientations for continued increase of the virtual source velocity of holes despite integration challenges.

Another observation is that devices with fully-silicided (FUSI) gate [85] have considerably higher carrier velocity than polysilicon gate devices with similar strain levels. These particular devices have lower halo doping to achieve the required threshold voltage and off current. So, these results deserve some more analysis to see whether higher virtual source velocity is in fact due to lower Coulomb scattering in these devices and if so, whether these results have any implication for the future devices with lightly doped ultrathin SOI or nanowire channels. For a given technology, carrier velocity increases with decreased electrostatic integrity [95] [8]. So, it is important to compare velocities at constant DIBL[9].

Figure 2-6 compares the estimated velocity of carriers in FUSI devices [85] with that of polysilicon gate transistors with similar strain level [12]. At a given DIBL, the virtual source velocity in FUSI devices is considerably higher than conventional MOSFETs, which highlights the fact that they suffer less from Coulomb scattering. However, the effective velocity, which determines the transistor performance, is quite similar in the two devices. This is mainly because FUSI devices have higher inversion capacitance and hence suffer more from the voltage drop across S/D series resistance according to (2.2). In other words, with less channel doping, these particular FUSI devices operate as a over-scaled polyisilicon-gate transistor. As will be discussed in Section 2.9, overscaling of the device, although it might lead to higher velocity or equivalently higher drive current, does not necessarily translate to higher performance as it will result in lower effective current. In addition to providing some insights into the implications of metal-gate transistors, this example gives an important message

[8]This is due to the fact that the critical length of scattering, l, decreases with increased DIBL.

[9]Care must be taken when comparing the carrier velocity in bulk and SOI transistors. The higher DIBL in SOI transistors is partly due to the floating body effect, and would not necessarily correspond to higher velocity.

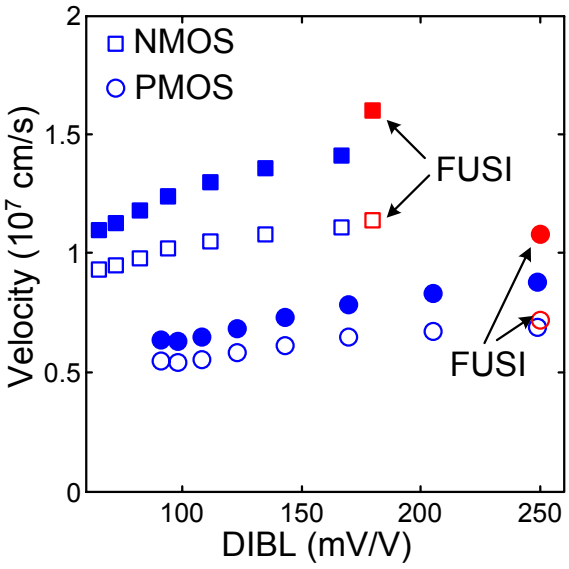


Figure 2-6: The extracted virtual source velocity (filled symbols) and effective velocity (open symbols) vs. DIBL for polysilicon [12] and FUSI [85] gate transistors. Lower doping in FUSI devices increases the ballistic efficiency and hence v_{x0} is higher. However, the effective velocity is comparable to that in polysilicon gate transistors due to higher C'_{inv}. Velocity estimation is done by assuming the following values for the C'_{inv}: 1.83 vs. 2.20 μF/cm^2 (NMOS) and 1.70 vs. 2.20 μF/cm^2 (PMOS) for polysilicon and FUSI gate transistors, respectively. Series resistance, R_S was assumed to be 80 vs. 85 $\Omega.\mu$m (NMOS) and 120 vs. 140 $\Omega.\mu$m (PMOS) for the two cases. Note that the extraction of the effective velocity, v, does not include any assumption for R_S.

about reducing the channel doping to alleviate the Coulomb scattering and hence enhance the transport properties: Unless short channel effects are controlled by means other than increasing C'_{inv}, such as thinning the transistor body, no benefit is gained from a lower channel doping.

Continuous increase of the virtual source velocity is needed in order for the commensurate scaling of the transistor performance with dimensional scaling to continue the historical trend presented in Figure 2-4 in the future technology nodes. Next chapter will study some of the materials and device structures that can be used to increase the carrier velocity. However, as will be discussed in the next section, unreasonably high velocities will be required beyond 45 nm node, unless dramatic change in the device scaling scenarios are adopted.

2.5 Future of Performance Scaling

When exploring the design space and opportunities offered by new device structures and material systems, most researchers rely on ITRS projections and usually find the combinations that meet the required I_{on} for the specified supplied voltage, I_{off}, electrostatic integrity, and parasitic components[10]. The required drive current in ITRS projections is calculated in order to fulfill the commensurate scaling of the intrinsic transistor delay based on the CV/I metric, which is too optimistic as highlighted in Section 2.2. Moreover, recent publications on 45-nm high-performance CMOS technologies [87, 89] show slow-down of the gate length scaling compared to ITRS projections. Also, ITRS requirements/assumptions for electrostatic integrity, effective oxide thickness, and parasitic components are too optimistic, as will be discussed in some detail in next section. Consequently, conclusions made based on such studies might not be applicable to real devices. Unfortunately, the underlying assumptions that these studies are based upon are often lost in the literature, causing more confusion. Extrapolatory circuit designs, though not affected by the choice of the delay metric, are still prone to the errors introduced by the unrealistic device parameters. It is thus imperative to explore the scaling trends of the future technology generations with more realistic assumptions than what is outlined in ITRS.

We use the analytical delay metric introduced in Section 2.2 with some assumptions about the device parameters to analyze the prospects of the performance scaling in the future "high performance" CMOS generations. A scaling scenario is first presented in the next section, followed by a discussion about major parasitic components. Outlooks of the performance scaling down to 15-nm HP node are studied and then we provide a case study of the 32-nm technology and examine the effectiveness of different approaches to meet the required delay.

[10]In its current version, ITRS does not explicitly specify any requirement on the electrostatic integrity, although models used to generate the tables are based on some optimistic values for subthreshold swing. Also, the requirements for the parasitic components inherent to the device, namely fringing capacitances and source/drain series resistance are optimistic as will be discussed in Section 2.7.

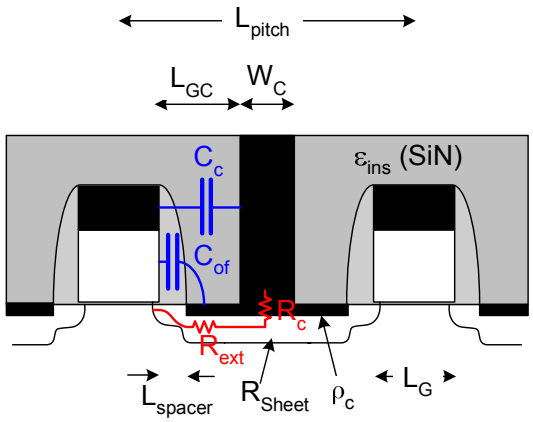


Figure 2-7: Illustration of the scaling device features and major parasitic components.

2.6 Device Scaling Scenario

We begin by observing that the key scaling parameter is the so-called contacted gate pitch, L_{pitch}, which historically has scaled by 0.7 from generation to generation and is desired to continue to do so. This will ensure doubling of the transistor count per each technology generation to reduce the cost and/or increase the chip functionality. Also, with the emergence of multi-core processors, it is more important to increase the number of available transistors than making them faster.

Figure 2-7 shows a schematic diagram of the transistor with the main parameters that define its structure and major parasitic components. Table 2.1 shows an aggressive scaling scenario for HP CMOS based on this pitch scaling assumption. Numbers for the past technology nodes are taken from the literature along with some personal judgments, while projected values for future generations are all based on judgments of what is likely technologically feasible, supported by some feedback from industry leaders.

Figure 2-8 illustrates the scaling trend of the key feature sizes according to Table 2.1. ITRS projections are also shown in dashed lines for comparison. The slow down of the gate length scaling is taken into account in our projection, while ITRS continues to scale the gate length with its "current" pace. We assume a gradual decrease in the inversion oxide thickness similar to ITRS projections. Although some

Table 2.1: High Performance (HP) NMOS Scaling Scenario

Generation	(nm)	250	180	130	90	65	45	32	22	15
Contacted Gate Pitch	L_{pitch}(nm)	910	636	445	310	220	155	110	78	55
Physical Gate Length	L_G(nm)	180	130	65	45	35	30	26	22	15
Inversion Oxide Thickness	T_{oxinv} (nm)	4.9	3.0	2.5	1.9	1.9	1.52	1.38	1.16	0.80
Physical Oxide Thickness	T_{oxphys} (nm)	4.5	2.5	2	1.2	1.2	4	3.4	2.5	1.2
Spacer Thickness	L_{spacer} (nm)	130	100	70	50	30	21	14	10	5
Gate to Contact Spacing	L_{GC} (nm)	240	162	125	87	60	40	26	20	15
Gate Height	T_G (nm)	250	250	160	140	90	80	65	55	40
Supply Voltage	V_{DD} (V)	1.8	1.5	1.4	1.2	1.2	1.0	0.9	0.8	0.7
DIBL	δ (mV/V)	70	80	100	120	150	150	150	150	150
Effective Subthreshold Swing	S^* (mV/dec)	100	105	110	110	120	120	120	120	120
Off Current	I_{off}/W (nA/μm)	0.5	3	10	40	100	200	300	300	300
Silicide Contact Resistance	$\rho_c(10^{-8}\Omega.\text{cm}^2)$	9	8	6	5	4	3	2.5	2	2
Sheet Resistance	$R_{\text{sheet}}(\Omega/\text{sq})$	300	250	200	200	200	200	200	200	200
Virtual Source Velocity	$v_{x0}(10^6\text{cm/s})$	8.1	8.6	9.2	10.8	13.8	16.5	16.5	16.5	16.5

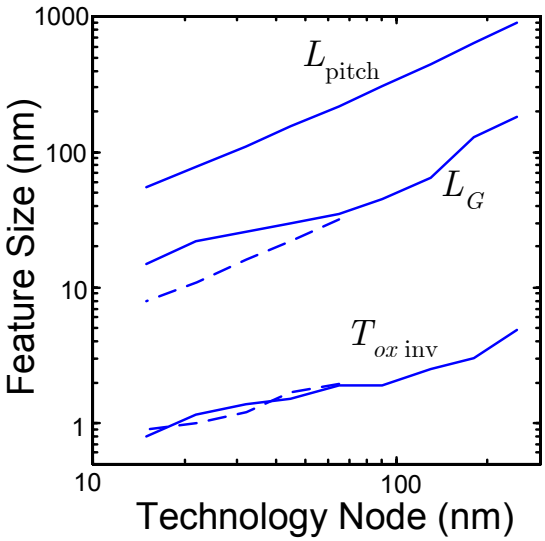


Figure 2-8: Scaling trend of key feature sizes for HP CMOS according to Table 2.1 (solid lines) compared with ITRS projections (dashed lines).

early publications on 45 nm CMOS continue to use silicon oxynitride as the gate dielectric [89], we assume a high-κ dielectric to allow a more aggressive scaling and be consistent with industry announcements.

Starting from 32 nm technology node, ITRS assumes that the off current decreases to about 100 nA/μm, whereas in our projection we let it increase to 300 nA/μm and stay there. Although ITRS does not specify the subthreshold slope and DIBL, from the specified threshold voltage and I_{off} values, the assumed subthreshold slope appears to be very optimistic, nearly ideal for double-gate structures. Also, note that our model to calculate the saturation threshold voltage based on the required I_{off} uses an effective subthreshold swing, defined by (2.3), which is slightly higher than the traditional subthreshold swing [11].

In addition, it is assumed that the virtual source velocity, v_{x0}, will increase somewhat from the 65 to the 45 nm technology and then stay constant, something that, at least for strain-engineered silicon is probably very optimistic, since the decreasing gate pitch will severely limit the space available for stressor materials [96].

[11] No attempt was made here to find the optimum threshold voltage and/or supply voltage for each technology node in order to minimize the total power-delay product or similar metrics. Such calculations require assumptions about the circuit activity which depends on the application of a specific circuit.

2.7 Scaling Trend of the Parasitic Components

Two models from the literature are employed to calculate the parasitic components, namely the parasitic capacitances and the series resistance. Parasitic capacitances are calculated using an analytical model [97]. The gate capacitance to contact vias is added to the "co-planar" value of C_f^*, with the added assumptions that the gate height is given by Table 2.1, the via half-pitch is equal to the via diameter, and the intervening-medium dielectric constant is equal to that of SiN. See Appendix A for details.

Figure 2-9(a) shows the scaling trend of the effective parasitic capacitance, C_f^*, as compared to the intrinsic gate capacitance, $C_{\text{inv}}L_G$. Shaded area represents the contribution of the parasitic capacitance between the gate electrode and source/drain contacts. Without this component, the calculated effective parasitic capacitance matches the value used in Section 2.2, i.e., 0.5 fF/μm. Note that the Miller effect at the drain side is taken into account when calculating the effective parasitic capacitance. It should be also noted that ITRS assumes that the contribution of the parasitic capacitances in the total load capacitance decreases from about 45% to less that 20% as the transistors are scaled. Possibilities for reducing the parasitic capacitance (or the $C_f^*/C_{\text{inv}}L_G$ ratio) are being sought [98, 99], as will be discussed in Section 2.9, yet numbers in the vicinity of those required by ITRS are very optimistic.

Although even in the current technologies, parasitic capacitance constitutes a major portion of the total load capacitance, its impact is even more significant. As shown in Figure 2-9(b), the parasitic capacitances are responsible for even a higher portion of the total switching charge and their contribution increases as the ratio V_{T0}/V_{DD} increases with scaling. This further contribution is neglected in the traditional CV/I metric used by ITRS.

The "transfer length" formula [100] is used for the calculation of the source/drain series resistance, R_S, based on some optimistic assumptions about sheet resistance (constant among generations), specific contact resistance (decreasing), and a constant

extension resistance $R_{\text{ext}} = 45\ \Omega.\mu\text{m}$.

$$R_S = R_{cs} + R_{\text{ext}} + \frac{\rho_c}{L_c}\coth(\frac{L_{con}}{L_c}) \tag{2.7}$$

with

$$L_c = \sqrt{\rho_c / R_{\text{sheet}}}, \tag{2.8}$$

where R_{cs} is the silicide contact resistance, L_c is the contact transfer length, ρ_c is the specific silicide contact resistance, and R_{sheet} is the diffusion sheet resistance underneath silicide layer. The last two parameters are given in Table 2.1. Our calculations show that R_S is likely to increase slightly with scaling despite the assumed reduction in specific contact resistance. This is mainly due to the fact that the area available for silicide formation decreases per Table 2.1. Again, ITRS requires that the series resistance decrease with gate length scaling. This has been almost constant for NMOS transistors over multiple technology generations and is unlikely to dramatically decrease. The advent of embedded SiGe in S/D has already cut the PMOS series resistance by about a factor of two, yet the corresponding numbers are still twice those of NFETs. Unless new silicide materials with lower barrier heights are introduced to decrease the specific contact resistance dramatically, it is improbable to achieve significantly lower numbers.

2.8 Prospects of Performance Scaling

Figure 2-10 shows the scaling trend of the intrinsic high-performance NMOS transistor delay, τ, calculated using (2.4) and based on the numbers given in Table 2.1. Experimental ring oscillator delay data for the past technology nodes [12] are also shown for comparison. The data demonstrate that up to 65-nm node, the calculated intrinsic delay has been scaled in proportion to the dimensional scaling in agreement with the experimental data. However, starting from 45-nm technology generation, the intrinsic delay stops decreasing and shows a counter-scaling thereafter. One might argue that this behavior is partly due to the fact that in the scaling scenario presented

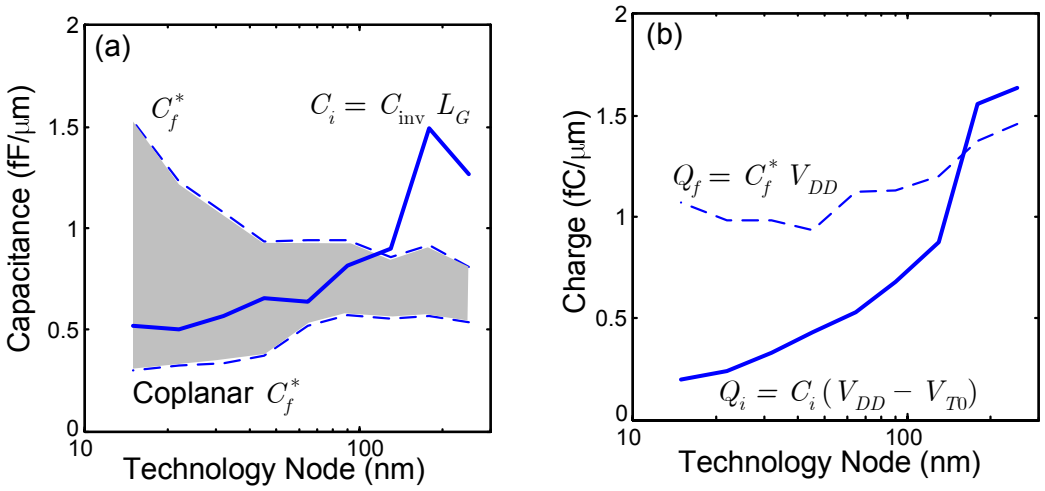


Figure 2-9: (a) Comparison of the effective parasitic capacitance, C_f^* , and the "intrinsic" inversion capacitance, $C_{\text{inv}}L_G$ as a function of the technology node. Shaded area represents the contribution of the parasitic capacitance between the gate electrode and S/D contact studs. Miller effect is taken into account for the drain side when calculating the effective parasitic capacitance. (b) Comparison of the parasitic and intrinsic charge vs. technology node. The effect of the parasitic capacitance is amplified as the ratio V_{T0}/V_{DD} increases with scaling.

in Table 2.1, we slowed down the gate length scaling. However, even a very aggressive scaling of L_G, such as what projected by ITRS does not change the trend reflected in Figure 2-10 considerably. In fact, as reflected in Figure 2-9, most of the switching charge is associated with the parasitic capacitances which do not scale very well with technology scaling. It should be noted that, a more aggressive gate length scaling does not necessarily translate to higher velocity as demonstrated in Figure 2-5.

To better understand the reasons behind the counter-scaling of the delay, Figure 2-11 shows the scaling trend of the virtual source velocity and the effective velocity. Although the parasitic resistance increases only slightly based on our assumptions, its impact on the effective velocity increases with increased C_{inv}. Hence, the effective velocity shows a drop starting from 45-nm technology, despite the fact that we kept the virtual source velocity constant.

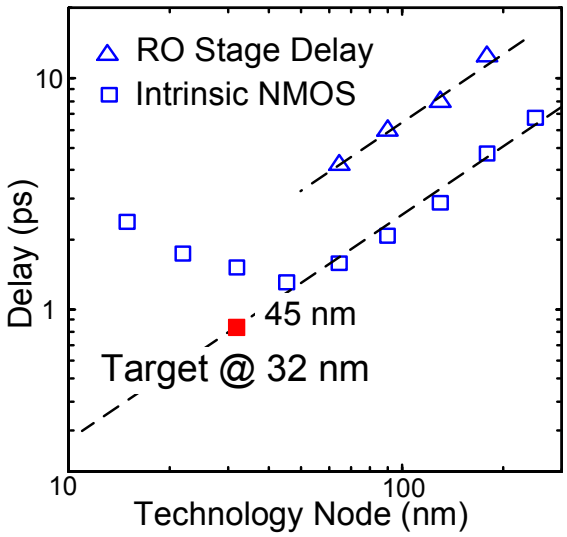


Figure 2-10: The calculated intrinsic NMOS delay vs. technology node based on the numbers in Table 2.1 (squares) and the experimental ring oscillator stage delay (triangles) [12]. Down to 45 nm node the transistor delay scales commensurate with the technology scaling. However, from 32 nm node onward the projected delay increases with device scaling. The required "target" delay at each future node is given by linear extrapolation of the historical data (in log-log scale).

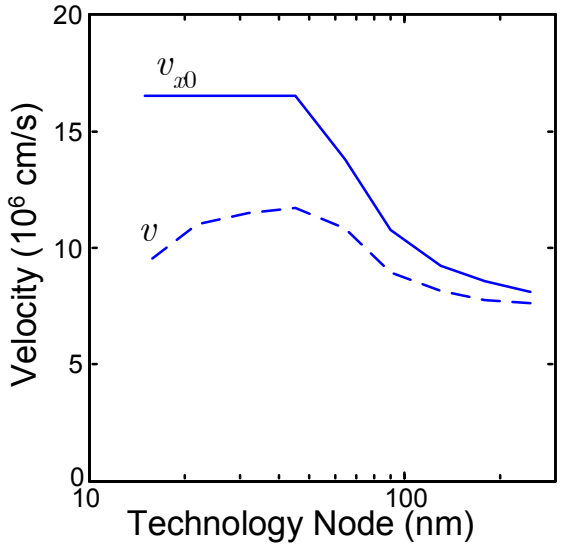


Figure 2-11: Scaling trend of the virtual source velocity, v_{x0}, and effective velocity, v. Despite the fact that the virtual source velocity is kept constant beyond 45-nm technology node and that the series resistance increases only slightly, the effective velocity drops because of the increase C'_{inv} according to (2.2).

2.9 Options for Commensurate Performance Scaling in 32-nm Node: A Case Study

The analytical delay expression presented in Section 2.2, allows us to easily explore the design space and identify major dependencies between device parameters and performance. Extrapolation of the historical data in Figure 2-10 provides the required delay at each technology node in the future in order to continue the historical trend [12]. In this section, we analyze different options that might allow commensurate scaling of the delay for NMOS transistors in the 32-nm high-performance technology. In doing so, we find it helpful to plot the required virtual source velocity as a function of the electrostatic integrity for each option. We believe this is an appropriate way of conveying the required transport properties, as the virtual source velocity increases with decreased electrostatic integrity according to Figure 2-6. The required v_{x0} can then be easily compared to what is feasible with a given material and/or structure. For easier illustration, we shall assume that the numerical values of the DIBL (in V/V) and the effective subthreshold swing (in V/dec) are identical, which is more or less true for $S > 80$ mV/dec. Below is an analysis of different options to achieve the target intrinsic delay of 0.8 ps at 32-nm node.

2.9.1 More Aggressive Gate Length Scaling

A more aggressive gate length scaling provides more space between the gate and S/D contacts and hence offers less parasitic capacitance. At the same time it leaves more space for the S/D silicide formation and possibly leads to less series resistance. Figure 2-12(a) compares the required virtual source velocity for the original design per Table 2.1 and a very aggressive scaling of the gate length to what is specified by ITRS. Although aggressive scaling of the gate length reduces the required velocity

[12] While we chose to plot the intrinsic transistor delay as a function of the gate length in Figure 2-4 to have a fair comparison and to show the basic trends, here we plot the delay vs. technology node to emphasize on what is expected from a new technology generation. One might argue that with the advent of multiple-core processors individual transistors are not necessarily required to operate faster. However, such argument can be made for any circuit-level design that improves the performance/lowers the power consumption.

considerably, the numbers are still beyond what is feasible by strained silicon. Yet, one should note that achieving the required electrostatic integrity with such an overscaled L_G is not easy. One potential approach is to use gate-all-around silicon nanowires to achieve the stringent electrostatic integrity requirements at such gate lengths. At the same time, since the channel can be kept essentially undoped without compromising electrotatics, one might expect velocities close to the ballistic limit though uniaxial strain would still be required. Of course, there are major technological challenges for implementing such devices, including packing the wires close to each other to deliver the required drive current without compromising either R_S or C_f^*.

2.9.2 Relaxed Gate Pitch

An alternative approach to provide more space between the gate and S/D contacts is to relax the pitch scaling. Of course, this comes at the expense of less packing density and contradicts Moore's law. However, for high-performance transistors this might be a feasible approach since they only constitute a small portion of the total transistor count. Also, with PMOS transistors approaching NFETs in terms of the drive current and performance, it is possible to recover some of the areal penalty associated with larger pitch by making PFETs narrower. Figure 2-12(b) illustrates the impact of a relaxed contacted gate pitch. Both of the above approaches have the added benefit of providing more room for applying strain to the channel and possibly obtaining higher velocities.

2.9.3 Thinner Gate Oxide

Increasing the inversion capacitance might seem a viable approach to increase the drive current and hereby decrease the delay. This can also be inferred from the delay expression of (2.4) where the ratio of the effective fringing capacitance to the intrinsic gate capacitance appears in the numerator. However, with more drive current, there would be more voltage drop across the source series resistance. In other words, the effective velocity which is the main lever for reducing the delay will be smaller with

increased inversion capacitance. Note that this analysis assumes no degradation of the interface quality due to the presence of a high-κ material. Yet, as reflected in Figure 2-12(c), increasing the inversion capacitance only marginally improves the performance. However, one might be able to exploit higher C_{inv} to better control the short channel effects and reduce the delay even further as shown schematically in Figure 2-12(c). Analyzing such options requires detailed assumptions about the exact device structure and doping levels and needs careful optimization of the gate workfunction. If the metal workfunction cannot be adjusted as desired, often the channel doping level needs to be reduced, opposing the short channel control offered by the higher C_{inv}.

2.9.4 Lower Parasitic Resistance

The above example hinted to the fact that to be effective, a increase in the C_{inv} should accompany a proportionate decrease in the source series resistance. This can be understood better from (2.2) where the $R_S C_{\text{inv}}$ product appears in the denominator to determine the effective velocity. However, as discussed earlier, reducing the series resistance is not easy to achieve. Figure 2-12(d) demonstrates that even radical reduction of the specific contact resistance by a factor of 2 ($R_S = 71\ \Omega.\mu\text{m}$) or complete removal of the contact resistance is not enough to bring the required virtual source velocity to values accessible by strained silicon.

2.9.5 Higher Power Dissipation

Increasing the supply voltage to 1 V or doubling the off current are not effective either, as shown in Figure 2-12(e). Both approaches marginally improve the performance for cases where short channel control is not very tight. This corresponds to the situations where the gate overdrive is less, and naturally an incremental increase in the overdrive by either decreasing the threshold voltage or increasing V_{DD} might be effective. Of course, in both approaches, the marginal improvement of the performance comes in the expense of higher power consumption.

2.9.6 Reduced Fringing Capacitance

Interestingly, this approach provides the best results as illustrated in Figure 2-12(f). This is of course not surprising as the parasitic charges constitute a major portion of the total switching charge, according to Figure 2-9(b). Despite technological challenges, this is also the most scalable approach as we proceed to the next technology nodes and it offers the added benefit of the reduced active power consumption.

As shown in Figure 2-12(f), even simple provisions to cut the gate height by a factor of 2, or using silicon dioxide instead of nitride in the spacers and as the intervening material between the gate and S/D contacts, are very effective in relaxing the requirements on the virtual source velocity. Of course, once silicon nitride is replaced by oxide, it might not be possible to assert the same amount of strain that the state-of-the-art MOSFETs are currently enjoying. However, it might be well feasible to achieve the required velocity with much less strain. Other approaches, such as reducing the polysilicon thickness or the number of contacts on the drain side (for wider transistors) have no or even positive impact on the amount of available strain.

2.10 Conclusions

This Chapter provided a review of the historical MOSFET performance scaling and studied basic dependencies of the performance on device parameters. Virtual source velocity of carriers was shown to be the main lever for increased performance over the past two decades. A roadmapping study was performed to determine the future of the performance scaling with some optimistic assumptions about device parameters. It was shown that uniaxially strained silicon, which is today's main channel material for high-performance transistors, is unable to meet the target performance. A case study for the 32-nm node illustrated that among the possibilities explored, reducing the fringing capacitance is the most promising approach to relax the velocity requirement. This is not surprising as the major part of the switching charge is due to the parasitic capacitances. Figure 2-12(f) shows that with reduced fringing capacitance

the required virtual source velocity is brought down to what can be achieved with strained silicon. Yet, one should note that it might be difficult, if not impossible, to achieve the required strain level without nitride stressors given that the available space for the stressor materials is also being reduced as the technology is scaled. For 22-nm node and beyond, new material systems will be required to extend the virtual source velocity above what is achievable with strained silicon. Next chapter reviews the basic dependencies of the virtual source velocity on material properties and explores some of the potential candidates to increase the carrier velocity.

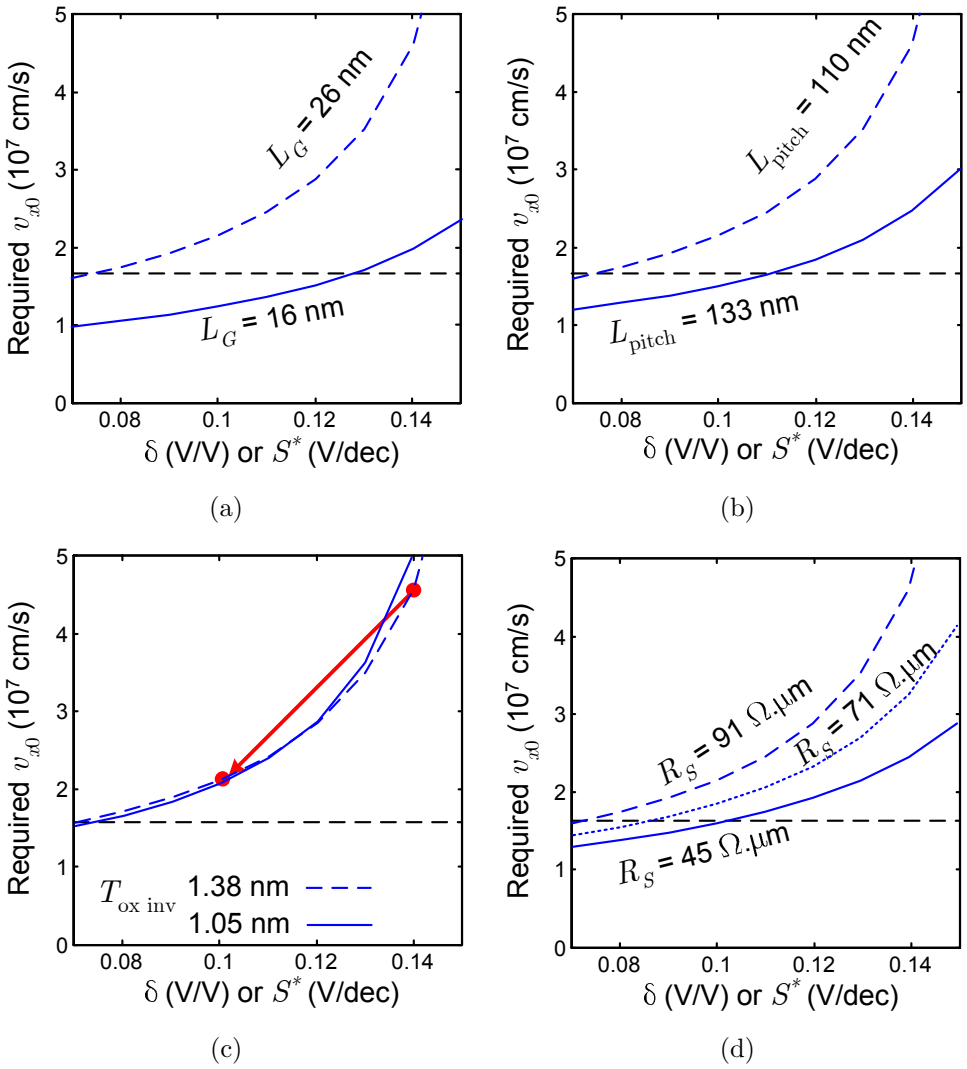
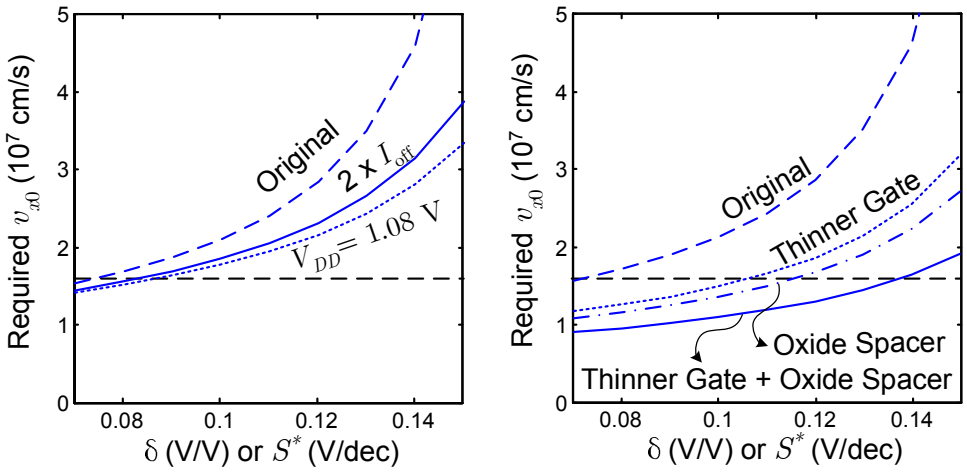


Figure 2-12 (continued on the next page.)



(e) (f)

Figure 2-12: The impact of different options on the required virtual source velocity to meet the target delay of 0.8 ps for the 32-nm high-performance NMOS as a function of the electrostatic integrity. The horizontal dashed line is the optimistically feasible v_{x0} per Table 2.1. (a) A more aggressive gate length scaling reduces the required velocity considerably. However, it is very challenging to control short channel effects. (b) A relaxed transistor pitch decreases the required velocity by reducing C_f^*. (c) Reducing the inversion oxide thickness is not very effective in relaxing the requirement on v_{x0}, due to voltage drop across source series resistance. It however opens a path to better control the short-channel effects. (d) Even dramatic reduction in the series resistance is not enough to bring the required velocity down to values feasible with strained silicon. (e) Either increasing the off current or increasing the supply voltage reduces the required velocity by a finite amount but at expense of higher power dissipation. (f) Reducing the fringing capacitance, by either using an oxide spacer ($\kappa = 3.9$) or reducing the gate height by a factor of two, appears to be very effective.

Chapter 3

Band Structure Engineering for Enhanced Transport

The analysis in Chapter 2 demonstrated that continuous increase in the carrier velocity is essential in order for the commensurate performance scaling to continue in future technology nodes. After introducing the basics of the transport in nanoscale MOSFETs, this chapter examines different approaches to enhance electron and hole transport by modifying the band structure. These approaches include biaxial and uniaxial tensile strain in silicon, III-V semiconductors, and germanium for enhanced electron transport and biaxial and uniaxial strain, different wafer and channel orientations, and germanium for enhanced hole transport. Prospects and limitations of each approach are analyzed based on simulations as well as experimental data from the literature wherever applicable.

3.1 Transport in Nanoscale MOSFETs

According to Lundstrom's scattering theory [50,101], MOSFET current is governed by the injection of carriers over the potential barrier in the channel (which is located close to the source when the transistor is in the saturation regime) as shown schematically in Figure 3-1. If the carriers encounter no scattering after they are injected into the channel (or more precisely, if they are not scattered back to the source), the "ballistic"

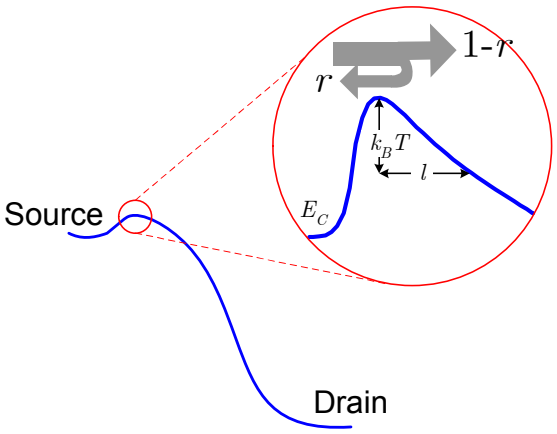


Figure 3-1: MOSFET current in saturation is governed by the injection of carriers over the potential barrier located at the source end of the channel. A fraction r of the carriers are backscattered to the source.

transistor current is given by:

$$I_D = Q_0 v_\theta, \tag{3.1}$$

where I_D is the transistor current, Q_0 is the inversion charge at the top of the barrier, and v_θ is the so called ballistic velocity of carriers, which in the non-degenerate limit is equal to the thermal velocity. This model further assumes that Q_0 is independent of the scattering inside the channel and the drain voltage (except for a finite shift in the threshold voltage due to DIBL). This assumption is supported by 2-D device simulations with a non-equilibrium Green's function (NEGF) modeling of the scattering [101]. So, in the saturation regime the inversion charge at the top of the barrier is simply given by $Q_0 = C_{\text{inv}}(V_{GS} - V_T)$, where V_T is the *saturation* threshold voltage.

In a real device, a fraction r of the carriers are scattered back to the source, as shown in Figure 3-1. Hence the effective velocity of carriers at the barrier, called virtual source velocity, is determined by this fraction [50, 101]:

$$v_{x0} = B v_\theta = \frac{1-r}{1+r} v_\theta, \tag{3.2}$$

where B is called the ballistic efficiency. Lundstrom's theory uses a third assumption that only scattering events that take place in the vicinity of the virtual source are

responsible for backscattering of carriers to the source[1]. Once a carrier passes the point where the potential has dropped by k_BT/q from the barrier top, the probability of return to the source is negligible. Hence, the reflection ratio, r, depends on the ratio between the backscattering mean free path of carriers, λ, and the distance over which the potential drops by the thermal voltage, the so-called critical length of scattering, l:

$$r = \frac{l}{l+\lambda}. \tag{3.3}$$

Therefore, the ballistic efficiency is given by:

$$B = \frac{\lambda}{\lambda + 2l}. \tag{3.4}$$

The validity of this latter assumption has been verified by means of Monte Carlo simulations [90, 91] and it is shown that the probability of backscattering to the source drops exponentially as the carriers move away from the barrier. So, a critical length of scattering, proportional to the distance over which the potential drops by the thermal voltage and an effective λ proportional to the actual mean free path can be defined for (3.4) to hold.

The above model is analogous to the Bethe condition for thermionic emission in a Schottky diode.

[1]In fact this latter assumption has been challenged by several Monte Carlo studies showing that scattering events all over the channel and even in the vicinity of the drain are responsible in determining the total current [102].

This apparent discrepancy can be addressed through the dependence of the critical length of scattering, l, on the scattering inside the channel. The velocity profile along the channel depends on the scattering mechanisms over the whole channel. For the current continuity to hold, the charge distribution also depends on scattering rates and so does the potential profile. In other words, accurate simulation of the electrostatics requires that the Poisson equation along the channel is solved self-consistently with the transport equations. With this procedure the distance over which the potential drops by k_BT/q in the vicinity of the barrier depends on the scattering rates in the entire channel.

3.2 How to Increase the Virtual Source Velocity?

Chapter 2 demonstrated that continuous increase of the virtual source velocity, v_{x0}, is essential to cancel the ever increasing effect of parasitic components in order to maintain the historical trend of performance scaling. In Section 3.1, above, its is seen that v_{x0} can be increased via increase of the backscattering mean free path, decrease of the critical length of scattering, or increase of the ballistic velocity. Analyses performed so far often focus on one of the above approaches and neglect the interplay of these three parameters. This section provides a theoretical background on how these parameters are determined in a hypothetical semiconductor. For the sake of simplicity, we will assume a single energy band with the effective mass approximation.

3.2.1 Ballistic Velocity

In a two-dimensional carrier gas, the carrier density is given by [103]:

$$N_{\text{inv}} = \frac{m_D}{\pi\hbar^2}\frac{k_B T}{2}\log\left(1+\exp(\frac{E_F-\epsilon}{k_B T})\right) = \frac{m_D}{\pi\hbar^2}\frac{k_B T}{2}\mathfrak{F}_0(\eta_F), \tag{3.5}$$

and the ballistic velocity is determined by [103]:

$$v_\theta = \sqrt{\frac{2k_B T m_y}{\pi m_D^2}}\frac{\mathfrak{F}_{1/2}(\eta_F)}{\mathfrak{F}_0(\eta_F)} \tag{3.6}$$

where m_x, m_y, and $m_D = \sqrt{m_x m_y}$ are the in-plane carrier effective mass parallel and perpendicular to the channel and the density of states mass, respectively, $\hbar$ is reduced Planck's constant, k_B is Boltzmann's constant, T is absolute temperature, E_F is the Fermi energy, ϵ is the minimum band energy, $\eta_F = (E_F - \epsilon)/k_B T$ is the reduced Fermi energy, and $\mathfrak{F}_n$ is the Fermi integral of the nth order.

In the non-degenerate limit, i.e. $\eta_F < 0$, the ballistic velocity reduces to the thermal velocity:

$$v_\theta = \sqrt{\frac{2k_B T}{\pi m_x}} \tag{3.7}$$

while in the degenerate limit, i.e. $\eta_F \gg 1$, it is approximately equal to

$$v_\theta = \sqrt{\frac{2k_BT}{\pi m_D^2}}\frac{2}{3}\eta_F^{1/2} \tag{3.8}$$

or

$$v_\theta = \frac{4\hbar}{3}m_x^{-3/4}m_y^{-1/4}\sqrt{N_{\text{inv}}} = \frac{4\hbar}{3\sqrt{m_C m_D}}\sqrt{N_{\text{inv}}}. \tag{3.9}$$

The above equation states that in the degenerate limit and at a given inversion charge density, the ballistic velocity is inversely proportional to the square root of the density of states mass, m_D, and the conduction mass, m_C. So, to increase the ballistic velocity in a MOSFET the structure/material or applied mechanical strain should be such that these two masses are reduced.

3.2.2 Backscattering Mean Free Path

It should be first emphasized that the mean free path that matters in determining the virtual source velocity is the backscattering mean free path which differs from the actual mean free path of carriers near the source; not all scattering events will cause a carrier to backscatter to the source. However, Monte Carlo simulations [90, 91] demonstrate that backscattering mean free path depends linearly on the actual mean free path of carriers near the source.

The notion of mobility in such short-channel transistors, where the scattering mean free path is comparable to the channel length, is a subject of controversy. However, a phenomenological mobility can always be extracted at low V_{DS}. Assuming that the mobility is constant across the channel and that the carriers at the top of the barrier are in a near-equilibrium condition, Rahman *et al.* [104] related this mobility to the backscattering mean free path by matching the low-V_{DS} drift-diffusion equation with the MOSFET scattering model:

$$\lambda = (\frac{2\mu}{v_\theta}\frac{k_BT}{q})\frac{\mathfrak{F}_0(\eta_F)}{\mathfrak{F}_{-1}(\eta_F)}. \tag{3.10}$$

This means that in order for the ballistic efficiency to increase, the low-field mobility,

μ, should be increased. Most of the attempts to date to increase the mobility by incorporating (mostly uniaxially) strained silicon in the channel are conceived based on this dependence. In fact, the traditional viewpoint only considers the decrease in the scattering rates upon applying mechanical strain, and hence [105]:

$$\frac{\partial v_{x0}}{v_{x0}} = (1 - B)\frac{\partial \mu}{\mu}. \tag{3.11}$$

It is also commonly believed that state-of-the-art MOSFETs operate at about 50% of their ballistic limit, i.e., $B \approx 0.5$ [47]. This means that any increase in the mobility will result in a 50% increase of the virtual source velocity or equivalently the drive current (in the absence of source series resistance) [105, 106].

Eq. (3.11) is valid only when the change in the mobility is purely a consequence of the change in the scattering rate and no change in the effective mass is involved. However, for all of the main scattering mechanisms in a two-dimensional inversion layer the mobility is inversely proportional to the transport and density-of-states masses [107]:

$$\mu \propto \frac{1}{m_C}\frac{1}{m_D}. \tag{3.12}$$

Again, a single sub-band and effective-mass approximation are assumed for simplicity. Compared with the effective-mass dependence of the ballistic velocity, Eq. (3.9)., where the ballistic velocity is inversely proportional to $\sqrt{m_C m_D}$, Eq. (3.12 suggests a power-law dependence between the ballistic velocity and mobility, i.e.,

$$v_\theta \propto \mu^{0.5}. \tag{3.13}$$

The above equation only holds when the change in the mobility is only a consequence of the modulation of the effective mass. In a general case where changes in the scattering rates are also involved,

$$v_\theta \propto \mu^\alpha \quad \text{where} \quad 0 \leq \alpha \leq 0.5. \tag{3.14}$$

3.2.3 Critical Length of Backscattering

The third parameter that contributes to the virtual source velocity is the critical length of scattering, l, often assumed to be equal (or proportional) to the distance over which the potential drops by the thermal voltage from the top of the energy barrier. The common belief is that this length is determined solely by the potential profile across the channel and hence it only depends on the drain voltage and the channel length [104]. Therefore, as the channel length is made shorter, so does l, and the transistor operates closer to the ballistic limit. However, the potential profile also depends on the scattering events through the channel [46, 90, 108]. With more velocity overshoot in the channel, the carrier distribution along the channel drops more abruptly. Accordingly, the electrostatic potential profile is modified to accommodate the change in the carrier distribution [108]. Self-consistent simulations to solve the electrostatics and transport together are thus needed to calculate the correct potential profile along the channel. Based on our non-equilibrium Green's function (NEGF) simulations using Nanomos [109] on a 30-nm transistor and with simple effective mass approximation to mimic the modulation of effective mass when uniaxial strain is applied along the channel [110, 111], l decreases with increasing mobility according to a power law, $l \propto \mu^{-\beta}$, where $\beta \approx 0.45$ [46] (Figure 3-2). The decrease in the l has also been observed in Monte Carlo simulations upon removing scattering mechanisms from the simulation, i.e., in ballistic transport [90]. According to our simulations, β decreases with shrinking the channel length in agreement with the observations in [90]. Also, β would be smaller when modulation of effective mass is less involved in determining the mobility enhancement, i.e., when $\alpha < 0.5$.

3.3 Effective Mass Considerations

The above discussion suggests that in order to maximize the virtual source velocity at a given inversion charge density, the device structure and channel material should be chosen in a way that minimizes both conduction and density-of-states masses. This argument, however, does not consider how the amount of the inversion charge

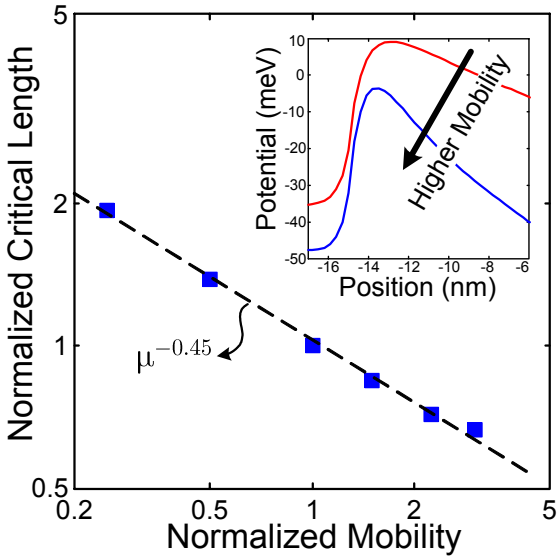


Figure 3-2: The critical length of backscattering, l, as a function of the mobility from non-equilibrium Green's function (NEGF) simulations (using Nanomos) of a 30 nm MOSFET. A power law relationship $l \propto \mu^{-\beta}$ with $\beta \approx 0.45$ is observed.

at a given supply voltage is affected by the choice of the effective mass. Since the inversion charge in saturation is given by $Q_{\text{inv}} = C_{\text{inv}}(V_{GS} - V_T)$, the question is how the inversion capacitance and the threshold voltage are affected by the choice of the effective mass.

3.3.1 Effective Mass Dependence of the Inversion Capacitance

The inversion capacitance in a hypothetical MOSFET is determined by a series combination of the oxide capacitance, C_{ox}, density-of-states capacitance, C_{DOS}, and quantization capacitance, C_{quant} [112]:

$$C_{\text{inv}}^{-1} = C_{ox}^{-1} + C_{DOS}^{-1} + C_{quant}^{-1}. \tag{3.15}$$

The density-of-states capacitance is given by [112]:

$$C_{DOS} = \frac{qQ_{\text{inv}}}{2k_BT}, \tag{3.16}$$

and its effect is negligible at high inversion charge density but could be significant in weak inversion. That is why 3-dimensional carrier gas and Boltzmann distribution are assumed in deriving 3.16 [112].

The quantization capacitance stems from the finite distance from the inversion layer centroid to the oxide interface. For bulk MOSFETs and assuming a triangular potential well at the semiconductor surface, this capacitance is given by [112]:

$$C_{quant} = \left(\frac{4\epsilon_s^2 q m_z}{9\hbar^2}\right)^{1/3} (Q_{\text{dep}} + \eta Q_{\text{inv}})^{1/3} \propto m_z^{1/3}\ E_{\text{eff}}^{1/3} \tag{3.17}$$

where ϵ_s is the semiconductor dielectric constant, Q_{dep} is the depletion charge density, E_{eff} is the effective electric field, and m_z is the effective mass normal to the interface. To maximize the inversion charge density at a fixed gate voltage overdrive, $V_{GS} - V_T$, a device structure that maximizes the quantization effective mass, m_z, is desired.

3.3.2 Effective Mass dependence of the Threshold Voltage

In order to maximize the gate overdrive at a given off-current, one needs to minimize the subthreshold swing, which in a bulk MOSFET is determined by the balance between the gate oxide capacitance, C_{ox}, the depletion capacitance, C_B, and a source and drain coupling capacitance that models 2-dimensional short channel effects (C_{SC} and C_{DC}), as shown in Figure 3-3.

$$S = \frac{k_B T}{q} \ln 10 \left(1 + \frac{C_B + C_{DC} + C_{SC}}{C_{\text{ox}}}\right) \tag{3.18}$$

In this sense neither the quantization nor the DOS effective masses are important in determining the subthreshold swing. However, the quantization effective mass affects the choice of the gate dielectric thickness through its impact on the gate tunneling current. With a larger quantization mass the tunneling probability decreases [113,114]

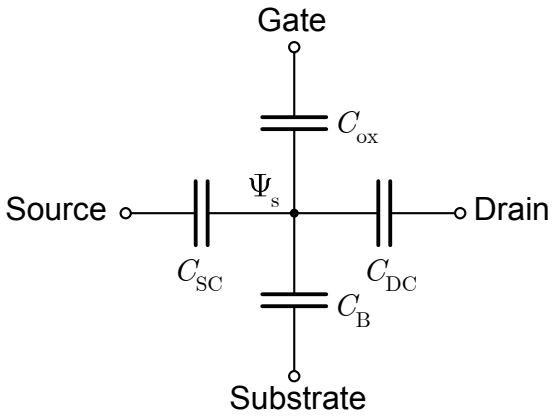


Figure 3-3: A simple model for the capacitive coupling between the MOSFET electrodes and the channel to determine the subthreshold swing: $S = k_B T/q \ln 10\,(1 + (C_B + C_{DC} + C_{SC})/C_{ox})$. In bulk MOSFET and assuming that the only coupling path between the source/drain and the channel is through the semiconductor, $C_{SC}+C_{DC} = \epsilon_s X_j/\gamma L_{eff}$, where X_j is the junction depth, ϵ_s is the semiconductor dielectric constant, L_{eff} is the effective channel length, and γ is a proportionality factor that depends on the details of the device structure and halo design. In a single gate SOI structure X_j should be replaced by the semiconductor thickness, whereas in double-gate SOI it should be substituted with one half of the thickness [115].

and hence for a given allowed gate leakage the oxide can be made thinner[2].

Furthermore, a higher quantization mass reduces the threshold voltage fluctuation due to variation in either channel doping (in bulk MOSFET), $\sigma_{V_T} \propto m_z^{-1/3}$, or the channel thickness (SOI), $\sigma_{V_T} \propto m_z^{-1}$. Figure 3-4 shows the threshold voltage fluctuation for 10% change in either channel doping or SOI thickness for different values of the quantization effective mass. Since the average off-current depends exponentially on the variations in threshold voltage [115], it is extremely important to minimize this variation.

The dependence of the off-current on the effective mass needs a more elaborate

[2]Under quantum confinement, the subband levels are moved to higher energies for carriers with smaller quantization effective mass. Hence, the band offset between the semiconductor and the dielectric will be smaller. For high-κ gate dielectrics that generally have smaller band offset compared to SiO_2, this could be significant. Also, since charge trapping to the defect levels inside the high-κ material is based on tunneling through the interfacial layer, higher quantization effective mass might also reduce the charge trapping. Although this might be benefecial in terms of threshold voltage instability of the devices, note that with the inversion charges closer to the dielectric interface the mobility might be degraded more.

analysis. In extremely short channel MOSFETs, in addition to the thermionic emission of the carrier over the potential barrier at the virtual source, tunneling through the barrier and band-to-band tunneling might contribute to the leakage current. The thermal velocity is given by (3.7) and is inversely proportional to the square root of the transport effective mass, similar to the dependence of the on-current. The threshold voltage depends logarithmically on the off-current, hence the increase in the off current due to decrease in m_C can be easily compensated by a small shift in the threshold voltage. The tunneling current, however, depends exponentially on m_C. Hence, a shift in the threshold voltage inversely proportional to m_C is required to compensate for the increase in the tunneling current. Depending on the severity of the tunneling current (which depends on the channel length), the subthreshold swing, and the available supply voltage[3], the required increase in the threshold voltage could be significant and a very small m_C might not be desirable.

The choice of the quantization effective mass affects the band-to-band tunneling. The effective bandgap increases significantly under quantum confinement and with a small quantization mass. This behavior has been exploited by several groups to reduce the band-to-band tunneling in small bandgap semiconductors such as Ge and InAs [117]. However, it should be noted that since the increase in the effective bandgap is proportional to $1/t_s^2$, fluctuations in the semiconductor thickness affect the tunneling current exponentially.

3.3.3 Effective Mass Dependence of the Mobility

Apart from small in-plane effective mass to increase the mobility, high quantization mass is needed to minimize mobility degradation due to confinement-enhanced phonon scattering [119] or thickness-fluctuation scattering [120] in ultra thin semiconductor channel structures. In the absence of any thickness-imposed confinement, the envelop function of carriers extends over some depth inside the semiconductor.

[3]Here is one example of how ITRS projections affect the conclusions made when exploring the material options. With a supply voltage of only 0.4 V, Rahman, *et al.*, concluded that (100) wafer orientation of germanium gives higher on current than (111) orientation for a given off current [116]. The conclusions could be quite different with a supply voltage of 0.7 V, which is more realistic

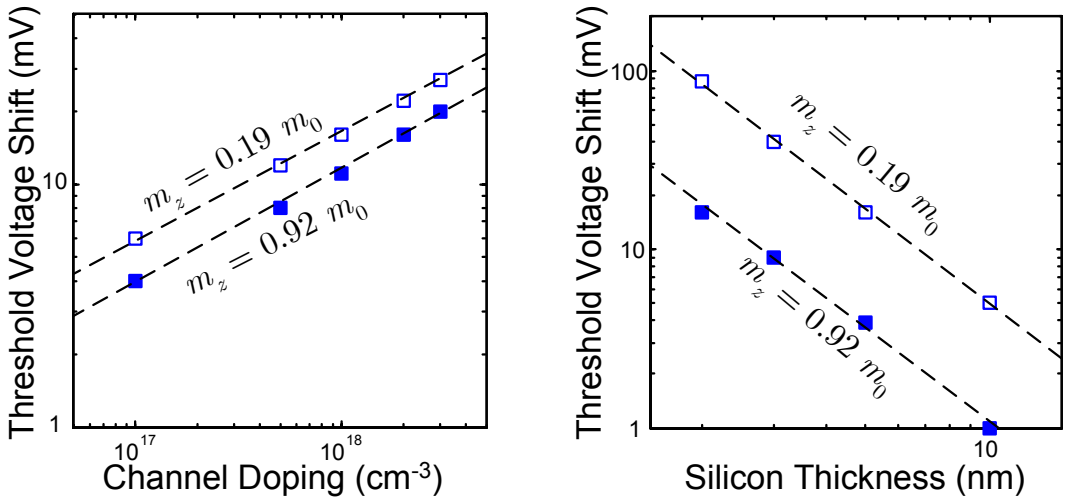


Figure 3-4: Threshold voltage fluctuation as a result of 10% change in either channel doping concentration (in bulk) or Si channel thickness for different values of the quantization effective mass. Oxide thickness is 1 nm in both cases and a double-gate structure with a doping of 10^{15} cm^{-3} was used for thin-Si channel case. Only 1-D effects were considered for the sake of simplicity through self-consistent simulations [118]. In short-channel transistors V_T fluctuation is even worse due to 2-D effects. Furthermore, discrete nature of the doping fluctuation or local thickness fluctuations lead to increased V_T fluctuation. Hence, the results presented here should be viewed as the lower limit on the threshold voltage fluctuation. In a bulk structure and assuming a triangular potential well, the quantum shift in the threshold voltage is equal to $\Delta V_T = \hbar^2/2m_z^{1/3}(9\pi q E_s/4\hbar^2)^{2/3}$, where $E_s = Q_{dep}/\epsilon_s$ is the electric field at the semiconductor surface. Hence the threshold voltage fluctuation is proportianl to $m_z^{-1/3}N_A^{1/3}$. For thin Si channel and using a particle-in-a-box model, $\Delta V_T = \hbar^2\pi^2/2m_z t_s^2$, where t_s is the semiconductor thickness.

Table 3.1: Guidelines for choosing the effective masses based on different requirements to increase performance.

Driver	m_C	m_D	m_z
Virtual Source Velocity	$\downarrow$	$\downarrow$	–
Mobility	$\downarrow$	$\downarrow$	$\uparrow$
Inversion Capacitance	–	$\uparrow$	$\uparrow$
Threshold Voltage	–	–	$\uparrow$
Tunneling*	$\uparrow$	–	$\downarrow$

* See the text for a discussion of the dependence of the direct and band-to-band tunneling on the effective mass and its implications.

The "width" of the inversion layer depends inversely on the quantization mass and the effective electric field, $W_{\text{inv}} \propto (m_z E_{\text{eff}})^{-1/3}$. Once the semiconductor thickness is reduced to a certain point, the inversion layer is squeezed by the extra confinement imposed by the thickness, $W_{\text{inv}} \propto t_s$. With this additional confinement is imposed, the phonon-limited mobility, which is proportional to the inversion layer width, drops below its value in bulk semiconductor. Hence, the thickness at which the mobility drops below its bulk value is roughly proportional to $(m_z E_{\text{eff}})^{-1/3}$.

The thickness fluctuation scattering can be modeled as a perturbation in width of a rectangular potential well. According to the Fermi Golden Rule, the scattering rate will be proportional to $\langle\psi|H|\psi\rangle^2$, where H is the perturbation Hamiltonian and ψ is the envelope function. Assuming a single subband, a fluctuation Δt_s in the semiconductor thickness will result in a potential fluctuation of $2\Delta t_s \hbar^2\pi^2/2m_z t_s^3$. Hence, the mobility limited by this scattering will be proportional to $m_z^2 t_s^6$. Again the thickness at which this extra scatting becomes significant is proportional to $m_z^{-1/3}$.

Table 3.1 summarizes these requirements. In the rest of this chapter this table is used as a guideline to select the proper choice of the wafer and channel orientation, material, and mechanical strain. This table suggests that a low conduction mass, m_C, moderate DOS mass, m_D, and high quantization mass, m_z, are desired.

3.4 Options for Enhanced Electron Transport

Review of the published work shows that applying uniaxial mechanical strain to the channel has already increased the electron virtual source velocity from about 10^7 cm/s in relaxed silicon to near 1.5×10^7 cm/s (see Figure 2-5) corresponding to about 60 % increase in the mobility. Yet, higher velocities are needed for continuous performance incraese as discussed in Section 2.8. This section briefly reviews the characteristics and limitations of the biaxial and uniaxial mechanical strain to enhance electron mobility and velocity in Si-based MOSFETs and then explores other material systems that potentially can be used to improve the transport properties even more.

3.4.1 Biaxial Tensile Strain

Despite the fact that biaxial tensile strain was a classical candidate to enhance electron mobility by almost a factor of 2, it was never used by the industry with the exception of a number research studies [21, 121, 122]. This is partly due to material integration challenges associated with epitaxial growth of strained Si and SiGe buffer layers. However, even if process issues are resolved, biaxially strained Si improves nMOSFET performance only slightly; 2× mobility enhancement translates to only 20-30 % increase in the drive current [21, 121, 122].

The mobility enhancement in biaxially strained Si is believed to be due to the band splitting between the conduction band minima. Upon applying biaxial strain, two of the six valleys in the Si conduction band are shifted to lower energies (Δ_2), whereas the other four valleys are shifted to higher energies (Δ_4). As a result, the interband optical phonon scattering is suppressed [107]. Although Δ_2 valleys have lower conduction and DOS effective masses, theoretically surface-roughness scattering is not affected very much unless a change in the surface roughness is assumed [123]; At high E_{eff} where surface roughness scattering is important, even in relaxed Si nearly 80% of the electrons are in the Δ_2 valleys. Most recent experimental work show that in fact there is reduction in the surface roughness for oxides grown on biaxially strained Si [124]. Although this is an interesting phenomenon, it is not clear whether high-κ

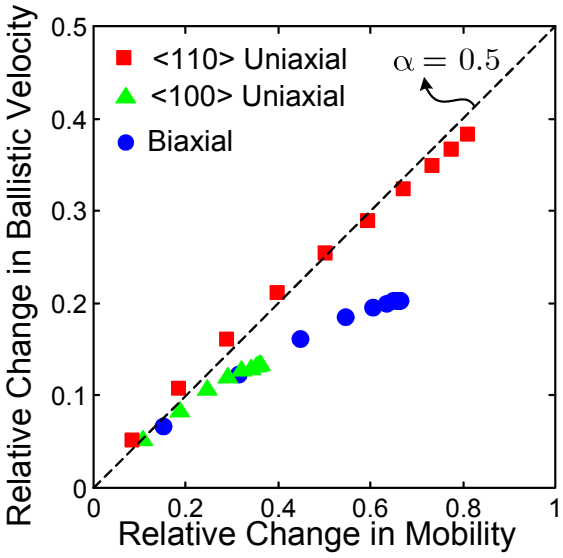


Figure 3-5: Relative change in electron ballistic velocity vs. relative change in mobility for different strains calculated based on the data in [110].

dielectrics needed for future technology nodes will benefit from a similar reduction in the surface roughness. Furthermore, Coulomb scattering which is important in determining the overal mobility in deeply scaled MOSFETs is not affected by biaxial tensile strain [125][4]. Finally, since the mobility enhancement is mostly governed by the reduction in the scattering rates rather than modulation of the effective mass, ballistic velocity is not enhanced as much, i.e., $\alpha < 0.5$. In fact, simulations show that for biaxial tensile strain $\alpha \approx 0.3$ as shown in Figure 3-5.

However, biaxial tensile strain might still be a viable approach for performance improvement:

1. For certain applications, such as pass-gate transistors, the linear transistor current, i.e., at low V_{DS}, is as important as the saturation drive current. Higher mobility is thus important even if velocity is not increased significantly.

2. Most high-performance chips operate at temperatures well above room temperature and will benefit more from suppressed phonon scattering.

3. Reduced optical phonon scattering leads to higher velocity overshoot near the

[4]Most recent experimental data suggest that Coulomb mobility limited by substrate doping is enhanced in strained silicon while mobility limited by interface states is degraded [126].

drain and hence results in smaller critical length of scattering as discussed in Section 3.2.3.

4. For ultra-thin silicon channel structures where confinement effects reduce the mobility through phonon scattering, biaxially strained Si extends its superior mobility to Si thickness of 3-4 nm [20, 127].

5. Biaxially strained silicon can be used as a starting material with either additional uniaxial strain [128] or to achieve higher levels of uniaxial strain by preferential relaxation of the strain [129][5].

3.4.2 Uniaxial Tensile Strain

So far, the most common approach to enhance electron transport has been the uniaxial strain in the [110] channel direction [130]. As far as the band splitting is concerned, this method is similar to the biaxial strain. However, an additional modulation of the effective mass in the Δ_2 valleys is involved. Simulations [110, 131] show that uniaxial tensile strain in the [110] channel direction reduces the conduction effective mass, while the in-plane effective mass perpendicular to the channel increases.

Figure 3-6 compares the change in the effective mass of electrons in the Δ_2 band in the direction parallel and perpendicular to a uniaxial strain in the [110] direction, calculated using a tight binding simulation [132] and empirical pseudopotential simulations reported in [131]. As seen in the figure, although there is agreement on the fact that electron effective mass in the direction of strain decreases with a tensile strain, there is no quantitative agreement. Comparison of pseudopotential models, a 30-band $k.p$, and *ab initio* simulations in [133] also show quite different results. Optimization of the simulation parameters are thus needed to reproduce the

[5]In fact based on the discussion in Chapter 1, showing the undesired contribution of the stress liners to the parasitic capacitances and given the fact that with scaled gate pitch there will be less space for stressors, this approach is very attractive especially for regular and long Si islands such as NMOS stripes in the SRAM cells.

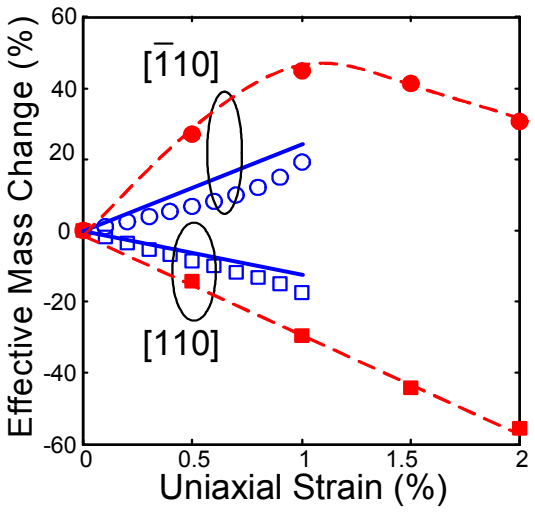


Figure 3-6: Relative change in the effective mass in the direction parallel and perpendicular to a uniaxial strain in the [110] direction. Open symbols are our tight binding simulation results, filled symbols are our empirical pseudopotential calculations, and solid lines are from empirical pseudopotential simulations reported in [131].

experimentally observed reduction in the effective mass[6].

One way to experimentally measure the modulation of the effective mass is to study the dependence of the mobility on external mechanical strain applied in the [110] direction on ultrathin SOI or SSDOI devices. In ultrathin SSDOI all of the electrons are in the Δ_2 band. The band splitting between Δ_2 and Δ_4 valleys is also large enough so that the interband phonon scattering is negligible. Hence, any change in the mobility with uniaxial mechanical strain [134] can only be attributed to a change in the effective mass. Of course, this does not clarify whether the change is in the conduction mass or the DOS mass. Nonetheless, experimental results [134] suggest that there should be 40% reduction in the effective mass (m_C and m_D combined) per 1% uniaxial strain in the [110] direction. Similar conclusion is made based on the measurements on ultrathin SOI transistors [110]. This should be compared to simulations performed in this work and those of [131], predicting a reduction of 25%

[6]In fact, the original version of the tight binding code used in this work gives very small change in the effective mass. The results presented in Figure3-6 are our first attempt to optimize the simulation parameters in order to reproduce experimental results. The earlier version of the code also lacked the modeling of the internal displacement, explained in Appendix B, and thus produced qualitatively wrong results.

and 15%, respectively.

By measuring the change in mobility for devices fabricated in different directions with respect to the strain direction, it is possible to decouple the change in the transport and DOS effective masses. Lauer [134] reported 40% increase and 35% decrease in mobility per 1% of uniaxial strain in the longitudinal and transverse directions, respectively. This gives 44% decrease and 36% increase for the effective mass in the tudinal and transverse directions, respectively, for each 1% of uniaxial strain. Rochette *et al.* [111] used a similar approach and concluded 46% decrease and 28% increase of the effective mass in these two directions, respectively. It should be noted that Rochette *et al.* performed the measurements on bulk MOSFETs where contribution from the Δ_4 band and possibly reduction in the interband scattering cannot be rulled out, whereas Lauer's measurements are on ultrathin SSDOI or ultrathin SOI devices, where a single Δ_2 band assumption is quite valid. These numbers can thus be used as a guide to calibrate band-structure models under uniaxial strain. Among the models presented in Figure 3-6 only our emprical pseudopotential calculations seem to provide relatively accurate results. The fact that none of the simulation methods are in quantitative agreement with experimental results, urges for better calibration of these models in order to explore the limits of electron mobility and velocity enhancement offered by uniaxial tensile strain[7].

Despite the uncertainty in the theoretical change in the effective mass, the fact that mobility enhancement is greatly due to the reduction in the effective mass is encouraging. Not only all scattering mechanisms are equally affected by the reduction in the effective mass, but also the ballistic velocity is significantly enhanced as shown in Figure 3-5.

[7]It appears that researchers are still relying on first-principle calculations to calibrate empirical methods such as tight-binding or empirical pseudopotential models. Using experimental data, at least for strain levels less than about 0.1% which are easily measured by mechanical bending, is a more reliable alternative.

3.4.3 Germanium

For the past two decades, germanium has been of interest as an alternative MOSFET channel material to improve transistor performance. Electron mobility in bulk Ge is nearly 2.7× that of silicon. However, despite earlier publications reporting very high electron and hole mobility in Ge MOSFETs [135–137], recent work to reproduce those results has been unsuccessful (see [138] for a comprehensive account of recent work). As will be discussed in details in the next chapter, this is mainly due to the presence of traps at the germanium-dielectric interface or inside the high-κ dielectric. This section, however, focuses on the intrinsic properties of Ge to see what can be expected from germanium-channel MOSFETs once integration challenges are resolved.

Unlike silicon, for which (100) wafer orientation is known to provide the best theoretical electron mobility and at the same time the best interface quality with SiO_2, options for wafer and channel orientation for Ge NFETs are still subject of discord. Figure 3-7 illustrates the conduction band valleys in Ge with different wafer orientations. Corresponding effective masses are collected in Table 3.2. In (100) Ge, all four valleys have the same small quantization mass and are equivalent. With (111) wafer, one valley has very large quantization mass and small conduction and DOS effective masses (L_1). The other three valleys have small quantization masses (L_3). Finally, in (110) germanium, two valleys have small m_z and the other two have moderate m_z. Furthermore, the conduction is not symmetric: with L_2 valleys preferentially populated, [110] channel orientation has the best transport properties. So, as far as transport properties are concerned, (111) wafer orientation is preferred, followed by (110) wafer.

$L - \Delta$ Interband Scattering in Germanium

Similar to most III-V semiconductors, germanium transport suffers from closeness of the satellite valleys to the conduction band minima. The Γ band is located only 0.14 eV above the L valleys, while the separation between Δ bands and L minima is reported to be between 0.15 and 0.21 eV [139, 140]. In bulk germanium this does not

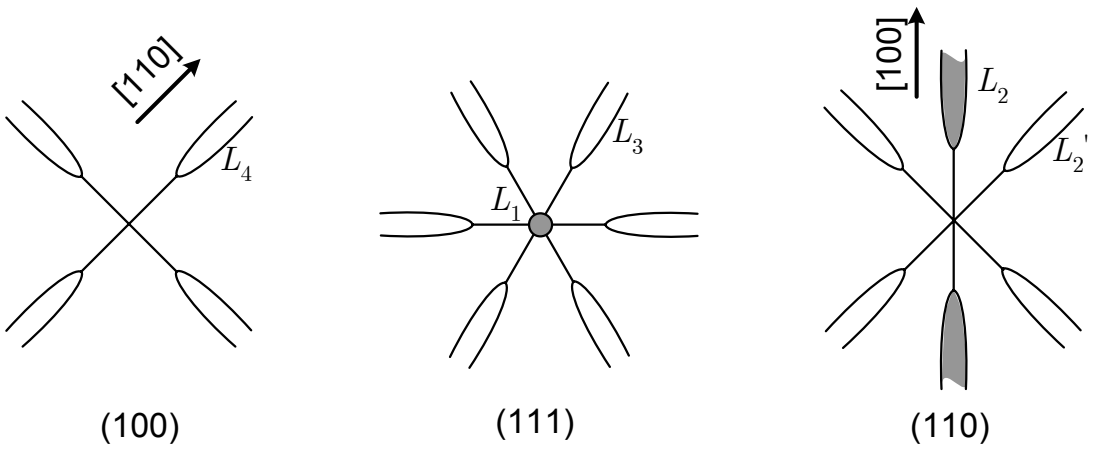


Figure 3-7: Germanium conduction band structure for different wafer orientations. In (100) Ge, all four valleys have the same small quantization mass. With (111) wafer, one valley has very large quantization mass and small conduction and DOS effective masses (L_1). The other three valleys have small quantization masses (L_3). Finally, in (110) germanium, two valleys have small m_z and the other two have moderate m_z. Furthermore, the conduction is not symmetric: with L_2 valleys preferably populated, [110] channel orientation has the best transport properties.

Table 3.2: Electron effective mass in the channel direction, m_x, normal to the channel, m_y, and normal to the wafer, m_z, and valley degeneracy, g in different wafer orientations of Ge. Effective masses are normalized to the free electron mass.

Wafer Orientation	Channel Orientation	valley	m_x	m_y	m_z	g	ΔE
(001)	[110]	L	0.15	0.15	0.12	4	
		Δ	0.20	0.20	0.95	2	0.15
			0.44	0.44	0.20	4	0.15
(111)	$[\bar{1}10]$	L	0.08	0.08	1.64	1	
			0.08	1.47	0.09	1	
			1.12	0.10	0.09	2	
		Δ	0.20	0.70	0.27	2	0.15
			0.58	0.24	0.27	4	0.15
(110)	$[\bar{1}10]$	L	0.08	0.58	0.22	2	
			0.22	0.12	0.08	2	
			0.20	0.95	0.20	2	0.15
			0.58	0.20	0.33	4	0.15

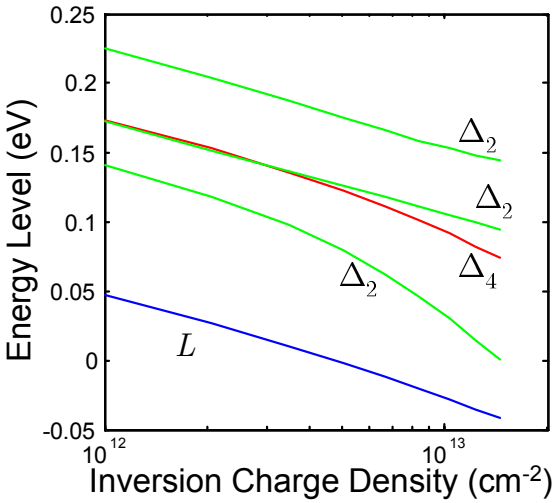


Figure 3-8: Electron subband energies in (100) germanium-on-insulator structure with a thickness of 5 nm and as a function of the inversion charge density.

affect the mobility as the energy of the optical phonons in germanium is only 35 meV. However, in Ge MOSFETs and under strong quantization, it is quite possible that the energy separation vanishes, leading to additional interband phonon scattering [141]. The Γ band has very small quantization mass and moves up very quickly once confinement is imposed to the carriers (either by increasing the gate voltage or in a GOI structure by thinning the germanium layer). Figure 3-8 plots the subband energies in (100) Ge with a thickness of 5 nm and as a function of the inversion charge density. As seen, since the L valleys have a very small quantization mass in this wafer orientation, they move up in energy under confinement so that Δ subbands contribute to the transport.

Degradation of mobility due to extra $L-\Delta$ interband scattering has been experimentally observed in bulk Ge samples under high hydrostatic pressure [142,143]. The L and Δ valleys have different dependences on the pressure: The L valleys move up in energy while the Δ bands move down. At very high pressures, all electrons are in Δ bands, making it possible to measure the electron mobility in these valleys. However, well before the crossover point, the mobility starts to degrade due to interband scattering. At the same time, measured Hall factor shows a significant increase above the ideal value of 1, which is a signature of multi-band transport (Figure 3-9).

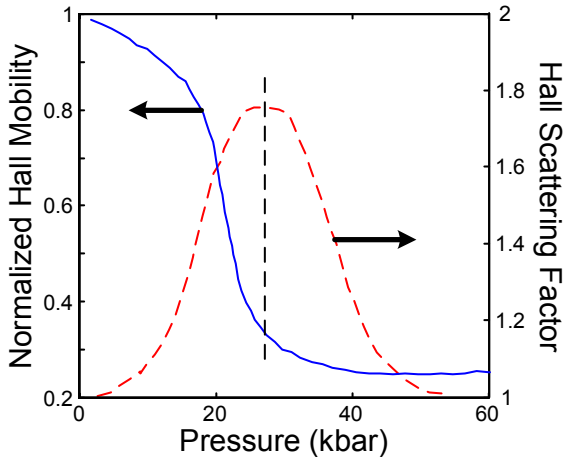


Figure 3-9: Normalized Hall mobility and scattering factor in bulk germanium under hydrostatic pressure [143]. As the pressure increases the L valleys go up as $dE_L/dP = 4.8 \pm 0.2 \times 10^{-6}$ eV/bar, while Δ valleys go down as $dE_\Delta/dP = -2.4 \pm 0.4 \times 10^{-6}$ eV/bar. So, band splitting between L and Δ valleys decreases as the pressure increases. At about 30 kbar the two bands cross. However, well before the cross-over point the mobility starts to decrease rapidly due to the additional $L - \Delta$ interband scattering.

Another area in which contribution of the Δ valleys to the transport and the additional interband scattering has been explored in the past is the high-field transport in germanium [141, 144]. In fact, it is shown that in a sub-micron Ge MOSFET and near the drain, roughly 20% of the electrons are in the Δ_2 bands [145], effectively reducing the velocity overshoot in that region. Hence, $L - \Delta$ interband scattering is quite deteriorating to the performance of germanium NMOSFETs and provisions should be sought to separate these two bands as much as possible.

Electron Mobility in Germanium

Figure 3-10(a) shows phonon-limited mobility, calculated in this work, for (100) germanium with different assumptions about the strength of the $L - \Delta$ interband scattering, i.e., deformation potential equal to 0, 4.4×10^8 eV/cm [146], and 15.0×10^8 eV/cm [147]. At higher inversion charge densities the mobility is degraded due to the additional interband scattering. Note that in these simulations it is assumed that the germanium is undoped and hence the effective electric field is not very strong, about 0.6 MV/cm at the highest N_{inv}.

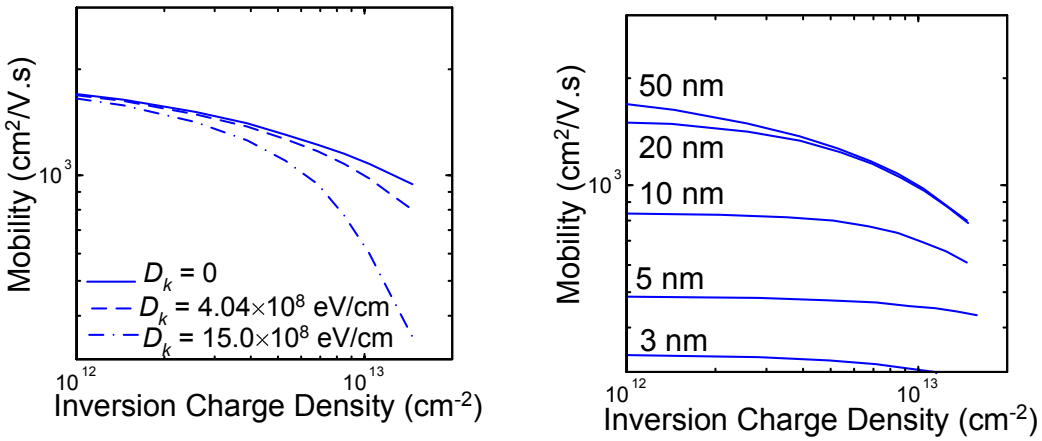


Figure 3-10: (a) Calculated phonon-limited electron mobility in bulk (100) germanium with different assumptions about the strength of the $L-\Delta$ interband scattering and as a function of the inversion charge density. (b) Phonon-limited electron mobility as a function of the germanium thickness in a germanium-on-insulator structure with (100) wafer orientation.

Figure 3-10(b) shows the calculations for different germanium thickness in a GOI structure and with $D_k = 4.04\times10^8$ eV/cm. Electron mobility is significantly degraded as the germanium layer is made thinner than about 10 nm[8]. This is in part due to increased phonon scattering due to quantum confinement [119] and in part due to additional interband scattering as the spacing between the L and Δ bands vanishes. In contrast, one of the valleys in (111) germanium has a very large quantization effective mass. Under quantum confinement, this valley is preferentially populated and there is virtually no $L-\Delta$ interband scattering. Furthermore, calculations show that there is little mobility degradation due to quantum confinement in ultrathin GOI with this wafer orientation [119], and even a mobility increase for germanium thickness in the 4-10 nm range and at low electron density.

[8]Note that these calculations are made with phonon scattering parameters fitted for the transport in bulk germanium. Scattering rates in the inversion layer could be quite different from bulk and fitting to experimental data is needed similar to what has been done for silicon [107]. The maximum electron mobility ever reported in germanium inversion layer is around 1000 cm^2/V.s [136], considerably lower than what the calculations in Figure 3-10 show.

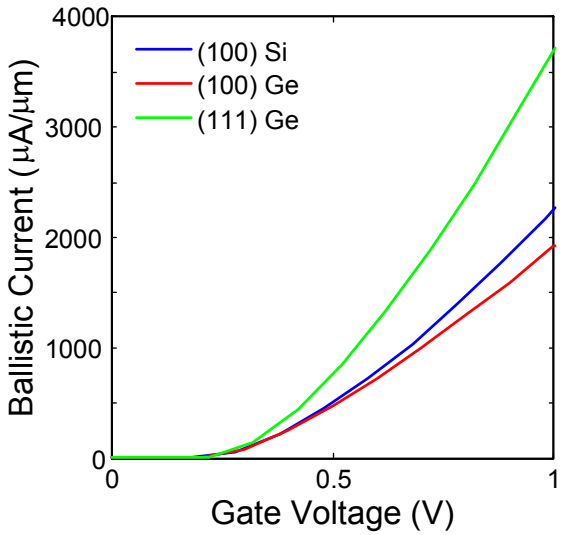


Figure 3-11: Comparison of ballistic current calculated for Ge with (100) and (111) wafer orientations and for (100) Si. With (111) orientation, Ge offers 80% improvement over silicon.

Ballistic Velocity in Germanium

Figure 3-11 compares the calculated ballistic current in germanium with different wafer orientations with that of silicon. With (100) wafer orientation, germanium does not offer any advantage over silicon, due in part to the smaller quantization effective mass which limits the inversion capacitance, and in part to the fact that the in-plane effective mass is not much smaller than in silicon. With (111) wafer orientation, germanium offers roughly 80% improvement over relaxed silicon, which is significant compared to state-of-the-art uniaxially strained MOSFETs. Whether or not applying mechanical strain improves electron velocity in (111) Ge requires simulations with well-calibrated band structure models or experimental demonstration.

3.4.4 III-V Semiconductors

Very high electron mobility in bulk III-V materials has fascinated many researchers for years to employ them as MOSFET channel material. In fact, their superior electron mobility is also experimentally observed in 2-D systems like MODFETs and MESFETs, and the main challenge for successful MOSFET demonstration is believed

to be gate dielectric formation with low density of interface states. However, from a theoretical point of view, two major challenges should be addressed before these materials are used for high-performance transistors:

1. The high electron mobility originates from the low effective mass of the Γ valley in these materials. However, the Γ band is symmetric, meaning that the DOS and quantization effective masses are also small, which are undesirable as discussed earlier. Furthermore, the low effective mass is only valid at the bottom of the conduction band. The non-parabolicity factor is typically large (and significant since DOS effective mass is small), meaning that at higher energies the equivalent effective mass is quite large. This becomes important either at high carrier concentrations or when the carriers gain energy as they travel along the channel. The latter becomes important at the drain-side of the channel where the non-parabolicity limits the velocity overshoot [145].

2. The separation between the Γ and the L and X band is usually not large. This, combined with the low quantization effective mass of electrons in the Γ band, means that under strong quantum confinement the satellite L and X bands will contribute to the electron transport. Not only do these bands have transport properties similar to the X and L bands in silicon and germanium, but also the additional interband scattering reduces the mobility further. Similarly, some of the high energy electrons near the drain are also transfered to the satellite valleys or have increased probability of having interband scattering [145][9].

Figure 3-12(a) and (b) illustrate the subband energies for a GaAs slab with thickness of 20 and 5 nm, respectively, and at an inversion charge density of 1×10^{13} cm^{-2}, while Figure 3-12(c) shows the electron distribution in that structure and the same bias condition. Self-consistent Schrödinger-Poisson simulations with parameters adjusted for GaAs and a gate oxide with EOT of 1 nm are used for these calculations.

[9]One might argue that even with these phenomena, III-V MODFETs and MESFETs in practice have significantly high transconductance and operate at very high frequencies. While this is true, we note that in such analog applications, transistor performance is quite different from digital application, discussed in Chapter 2: the inversion charge is usually small and transistors are relatively large to drive the required current.

Only the first 4 subbands are shown for clarity and the energies are referenced to the Fermi level. In the thicker GaAs layer, the first three subbands are Γ valleys, whereas in the thinner layer both L and X ladders contribute significantly to the inversion charge.

Figure 3-13 shows simulated inversion $C-V$ characteristics of a GaAs structure with thickness of 20 and 5 nm and with a physical oxide thickness of 1 nm. Due to very small quantization effective mass, the inversion capacitance is very small, nearly 1/3 of C_{ox}. This limits the available inversion charge at a given gate overdrive as discussed earlier. At electron densities higher than about 7×10^{12} cm^{-2}, the satellite valleys contribute to the inversion charge and hence there is a kink in the $C-V$ characteristics[10]

In semiconductors with small bandgap, such as InSb and InAs, the band spacing between the Γ band and the satellite bands is quite high. However, the small bandgap translates to higher leakage current (mostly band-to-band tunneling). Ternary alloys, such as InGaAs, provide acceptable bandgap, enough band separation to satellite bands, and significantly high mobility. Yet, small quantization effective mass still limits the inversion capacitance and hence the available inversion charge. Table 3.4 gives a tentative comparison of III-V and strained Si MOSFETs. It can be clearly inferred from this table that smaller inversion charge density in the III-V devices cancels out the benefit offered by their higher carrier velocity. Note that since the switching charge is dominated by the parasitic capacitances, there is little benefit in reducing the inversion capacitance.

3.5 Options for Enhanced Hole Transport

The valence band structure in most semiconductors is highly warped. The $E-k$ dispersion relation is anisotropic and the non-parabolicity is high. Although accurate treatment of the band structure in the inversion layer requires self-consistent simu-

[10]The band non-parabolicity is not considered in these simulations. Inclusion of the non-parabolicity results in higher density-of-states for the Γ band and hence the contribution of the satellite valleys will be less. However, non-parabolicity itself degrades the transport properties.

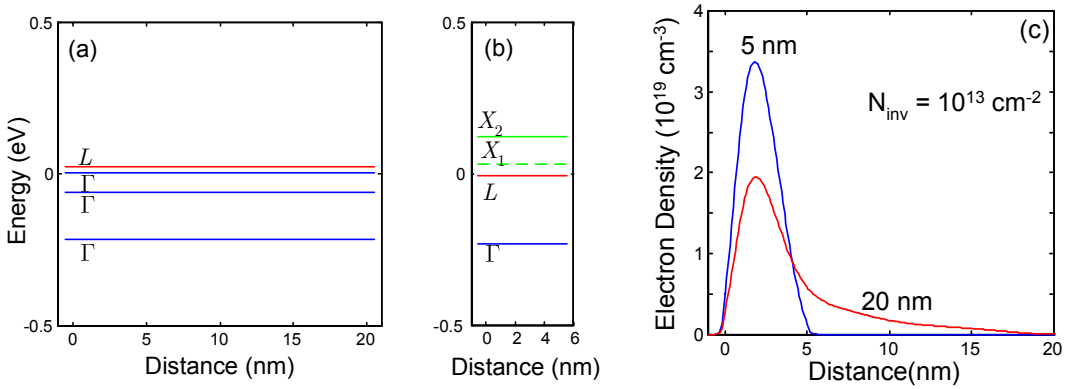


Figure 3-12: Subband energy levels in a GaAs slab at an inversion charge density of 10^{13} cm^{-2} and with a thickness of (a) 20 nm or (b) 5 nm. Only the first 4 subbands are shown for clarity. (c) Electron distribution in the two cases.

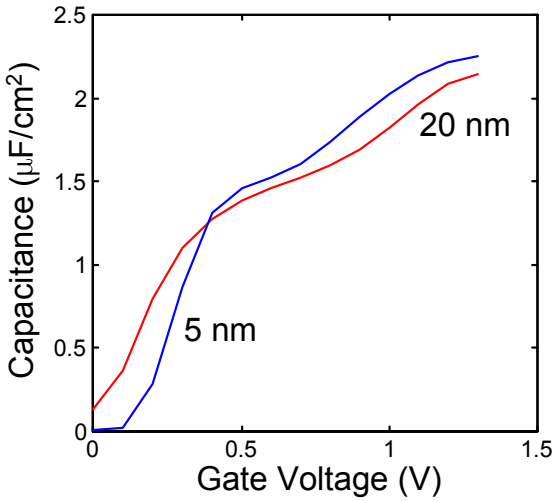


Figure 3-13: Simulated $C-V$ characteristics for a GaAs with a thickness of 20 or 5 nm and with physical oxide thickness of 1 nm. There is significant shift to the right as the slab thickness is reduced due to quantum confinement. Also, since the quantization effective mass in the Γ valley is very small, the inversion capacitance is much smaller than the physical oxide capacitance. At higher electron densities satellite valleys contribute and the $C-V$ characteristics show a kink.

Table 3.3: Properties of direct band-gap III-V semiconductors: bulk electron mobility, effective mass and non-parabolicity in the Γ band, bandgap, and the separation between the Γ and the L and X band.

Material	μ_e (cm^2/V.s)	m_e (m$_0$)	α (eV^{-1})	E_g (eV)	$\Delta_{\Gamma-L}$ (eV)	$\Delta_{\Gamma-X}$ (eV)
GaAs	9.4×10^3	0.063	1.16	1.42	0.32	0.45
InP	5.0×10^3	0.082	0.61	1.34	0.59	0.85
GaSb	3.0×10^3	0.041	-	0.73	0.08	0.31
InAs	4.0×10^4	0.023	1.4	0.35	0.73	1.02
InSb	8.0×10^4	0.014	4.1	0.17	0.51	0.83
$In_{0.53}Ga_{0.47}As$	8.0×10^3	0.041	-	0.80	0.60	0.70

Table 3.4: Comparison of the performance in a hypothetical III-V MOSFET with a typical uniaxially strained Si transistor. Common parameters are $T_{\text{ox}} = 0.7$ nm, $V_T = 0.4$ V, $L_G = 15$ nm, $\delta = 150$ mV/V, and $R_S = 80\Omega.\mu$m. Smaller inversion capacitance of the III-V cancels out the benefit offered by higher electron velocity.

Parameter	III-V	III-V	Strained Si
V_{DD} (V)	0.6	0.8	0.8
T_{oxinv} (nm)	2.0	2.0	1.0
C_{inv} (μF/cm^2)	1.72	1.72	3.45
Q_{inv} (μC/cm^2)	0.34	0.69	1.38
v_{x0} (10^7 cm/s)	4.0	3.0	1.6
v (10^7 cm/s)	2.3	2.0	1.0
I_{Dsat} (μA/μm)	800	1380	1380
I_{eff} (μA/μm)	109	585	585
τ (ps)	8.50	2.2	2.3

lations to account for the change in the band structure under quantum confinement, qualitative conclusions can be drawn using bulk band structure. This section reviews some of the options for enhanced hole transport.

In bulk semiconductors, the valence band can be described in terms of the degenerate light-hole (LH) and heavy-hole (HH) bands and the spin-orbit (SO) band separated from the other two by Δ_{so}. Light holes have a smaller quantization mass and move to higher energies under quantum confinement. Also, the DOS effective mass of LH is relatively small. Hence the main contributor to the hole transport in an inversion layer is the HH band. Of course, under quantum confinement, the subbands are no longer LH and HH. More appropriately they can be named as HH-like and LH-like. In the following, hole transport is qualitatively analyzed based on the shape of the HH under different conditions. Numerical results from $sp^3s^*d^5$ simulations are presented for illustrative purposes or to show band-structure dependence on the semiconductor thickness. We begin by analyzing different options to enhance hole transport in silicon, namely alternative channel and wafer orientations and different strain configurations and then discuss the benefits and limitations of using germanium as the channel material.

3.5.1 [100] Channel Orientation

As discussed in Appendix C, on a (100) wafer the smallest effective mass of HH is seen in the [100] direction. So, PMOS transistors with the channel in [100] direction can benefit somewhat from a smaller conduction mass. Of course, the DOS effective mass is still obtained by integration over all directions on the (100) plane and does not change with the choice of channel direction. Since electron transport on the (100) wafer is isotropic, the whole layout can be simply rotated 45° with respect to the major flat or the notch on the wafer. Nearly 20% mobility enhancement has been reported in SiGe-channel MOSFETs by making the channel in the [100] direction [26]. This is much less than typical mobility enhancement that is currently achieved by applying uniaxial mechanical strain on the transistors fabricated with the conventional [110] channel orientation. Nevertheless, for cases that only biaxial (compressive) strain is

available or for future thechnologies if it was determined that there is not enough room to apply the uniaxial mechanical strain, this approach may come to the rescue. Of course, provisions should be sought to optimize process steps such as implantation and annealing since these steps might depend on the crystalline orientation.

3.5.2 (110) and (111) Wafer Orientations

According to Table C.3, the effective mass in the [100] direction is the smallest. Hence, with a (100) wafer, the quantization effective mass will be relatively small, which is not desired. Higher quantization effective masses can be accessed if the transistor is made on a (111) or (110) wafer. Furthermore, with these wafer orientations it is possible to obtain much higher mobility and velocities. Nearly 2.8× and 1.8× hole mobility enhancement has been measured in long channel transistors fabricated on (110) wafers with channel in the [110] and [100] directions, respectively [28]. The mobility enhancement with a [211] channel on (111) wafers is roughly 60% [28]. With a V-groove cut on a (100) wafer by means of anisotropic wet etch to access (111) planes, 60% mobility enhancement has been reported with the channel in the [110] direction [148]. One major challenge for adopting (110) or (111) wafer orientations for PFETs is that these orientations have the lowest electron mobility [28,149]. Different schemes to integrate (110) PFETs with (100) NFETs by a mixture of preferential epitaxial growth and wafer bonding has been proposed [28].

Mobility enhancement achieved with these alternative wafer orientations is in part due to increase in the quantization effective mass of heavy holes. With quantum confinement the band splitting between the LH and HH increases due to their m_z mismatch in a (110) and (111) direction. This does not lead to an increase in the ballistic velocity. No wonder that 2.5× mobility enhancement does not translate to the same amount of increase in the virtual source velocity [28].

3.5.3 Uniaxial Compressive Strain

Uniaxialy compressive strain exerted by means of compressive stressors or embedded SiGe S/D, has been very successful in increasing hole mobility in the past few years. Hole mobility enhancement of nearly 4× [89] has been achieved by tailoring the strain distribution on (100) wafers. The heavy hole valence band is highly warped, with the highest effective mass in the [110] and the lowest in the [100] direction. A uniaxial compressive strain in the [110] direction reduces the effective mass in the [110] direction, which is the usual channel orientation. Figure 3-14 shows the contours of constant energy as a function of the in-plane wavenumber and under different uniaxial strains. A small uniaxial strain reduces the effective mass at the top of the valence band and in the [110] direction significantly. However, the dispersion relation at higher k values is nearly unchanged. Further increase of the strain, reduces the equivalent effective mass at higher k values.

At a given hole density, the dispersion relation only up to the point where the band is occupied by the holes is important in determining the ballistic velocity. Assuming a single subband structure, this would be the Fermi wavenumber, k_F. However, mobility depends on the shape of the band structure up to energies $\hbar\omega_{op}$ higher, where $\hbar\omega_{op}$ is the energy of the optical phonons. Based on the observations in Figure 3-14, while mobility continues to benefit from the change in the dispersion relation at high energies (and possibly reduced interband scattering), velocity enhancement slows down once a certain level of uniaxial strain is reached. Of course, this level of strain depends on the hole density and the amount of quantum confinement. At higher hole density the band is occupied up to higher energies, and hence the critical strain level will be higher. With higher confinement, the band structure will be more asymmetric, which means that higher strain levels are needed to remove the [110] wing.

Figure 3-15 shows the relative change in the ballistic velocity as a function of the relative change in the mobility based on the simulations in [150]. For hole density of 10^{13} cm^{-3} and up to a uniaxial strain of 0.5%, the ratio $(\partial v_\theta/v_\theta)/(\partial\mu/\mu)$ is about 0.5,

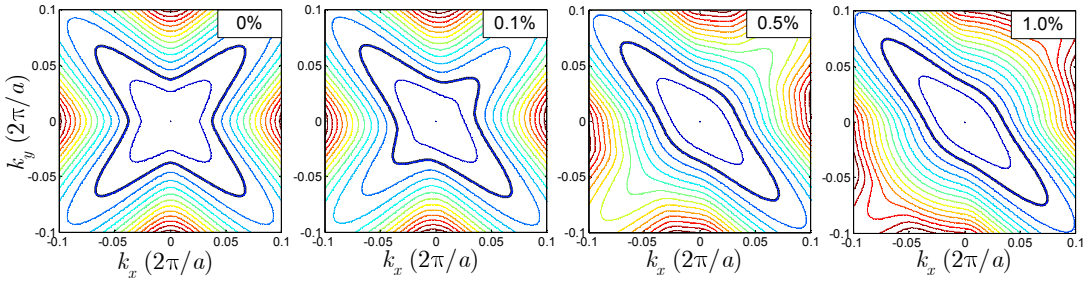


Figure 3-14: Contours of the constant energy as a function of the in-plane wavenumber in relaxed bulk Si and under different levels of uniaxial compressive strain in the [110] channel direction. The spacing between the contours is 0.25 meV and simulations are done using a tight binding code [132].

comparable to the dependency for electrons under uniaxial [110] strain. At higher strain levels, the ratio drops to about 0.3 at 1% uniaxial strain. The slow-down of velocity enhancement sets a limit on the performance improvement that can be achieved by uniaxial compressive strain. However, this technique has certain benefits: The mobility enhancement (at least at low strain levels) is caused by the modulation of the effective mass. Hence, all scattering mechanisms benefit from reduction of the effective mass in the [110] channel direction. The top valence band is still a heavy hole band. So, there is no competition between the strain and the quantum confinement. This is not the case in biaxial tensile strain, discussed in the next section. Furthermore, as shown in Figure 3-16, the reduction in the effective mass is maintained in ultrathin SOI, where in relaxed Si band structure the asymmetry is very strong and the effective mass in the [110] direction is very large.

3.5.4 Biaxial Tensile Strain

Similar to the case for electrons, biaxial tensile strain in the (100) plane, was traditionally believed to increase hole mobility significantly. In fact in bulk silicon, applying a biaxial tensile strain not only lifts the degeneracy between light (LH) and heavy (HH) hole bands, thereby reducing phonon scattering, but also preferably populates LH, reducing the effective mass considerably. More than a factor of 2 increase in the

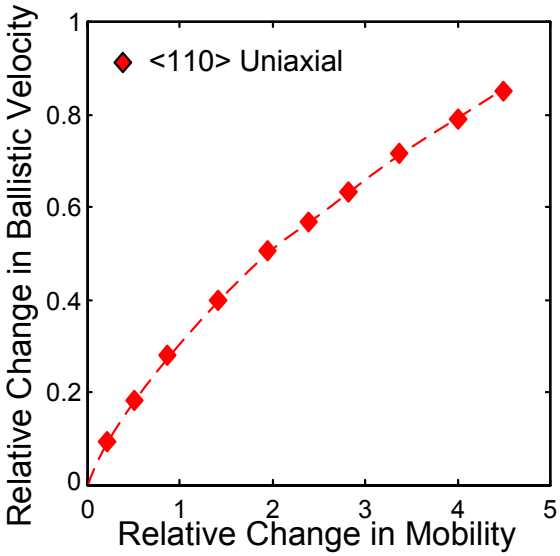


Figure 3-15: The relative change in the ballistic hole velocity as a function of the relative change in the mobility in bulk Si under uniaxial compressive strain in the [110] direction and at a hole density of 10^{13} cm^{-3} based on simulations in [150].

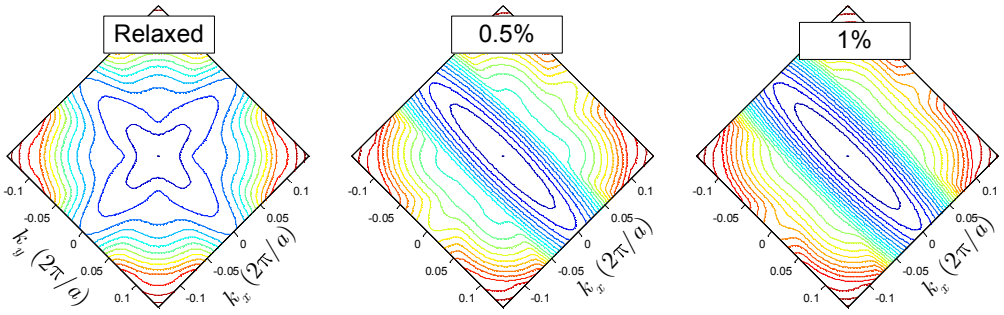


Figure 3-16: Contours of the constant energy as a function of the in-plane wavenumber in relaxed ultrathin Si with 3 nm thickness and under different levels of uniaxial compressive strain in [110] channel direction.

mobility is thus expected. However, this strain configuration has serious shortcomings if used for MOSFETs.

In the effective mass approximation, the quantization mass for light and heavy holes is 0.20 m_0 and 0.29 m_0, respectively. So, in a potential well LH subbands move to higher energy levels, contradicting the band splitting caused by the biaxial strain. As a result, while there is some benefit because of reduced phonon scattering at lower effective electric fields, mobility enhancement disappears at high E_{eff} [121]. In fact, since confinement in the potential well causes the LH and HH to split (of course with HH preferably populated), a small level of biaxial strain might cancel this band splitting and reduce the mobility [121]. This anomalous behavior has also been observed by applying mechanical biaxial strain to either relaxed or biaxially strained wafers [151]: In relaxed Si, hole mobility degrades by applying biaxial strain, whereas in strained Si mobility increases. High levels of strain are thus needed to maintain the band splitting at high E_{eff}.

A similar phenomenon is observed in ultra-thin SOI. For silicon films thinner than about 10 nm, the degeneracy of the light and heavy holes is already lifted due to the strong confinement. The dispersion relations are also quite different from those in bulk [152, 153]. Whether or not biaxial strain provides any benefit depends on how the combination of the carrier confinement and the applied strain affect the band structure. Recent experimental data [154] illustrate a complicated picture similar to bulk: while there is no mobility enhancement from 1.25% of biaxial tensile strain in a 3 nm strained Si on insulator, increasing the tensile strain to 1.67% improves the mobility by about 25%.

Figure 3-17 shows the effect of biaxial strain on the valence band energy levels in bulk silicon and ultrathin body SOI. In UTB SOI the degeneracy of the valence band at the Γ point is already lifted due to the strong confinement. In contrast to bulk silicon, where a biaxial tensile strain splits the HH and LH bands, in UTB SOI and for moderate values of tensile strain, these bands are actually converging. This is due to the fact that the biaxial tensile strain and confinement work in the opposite directions. Once a certain strain level is reached, the HH and LH bands

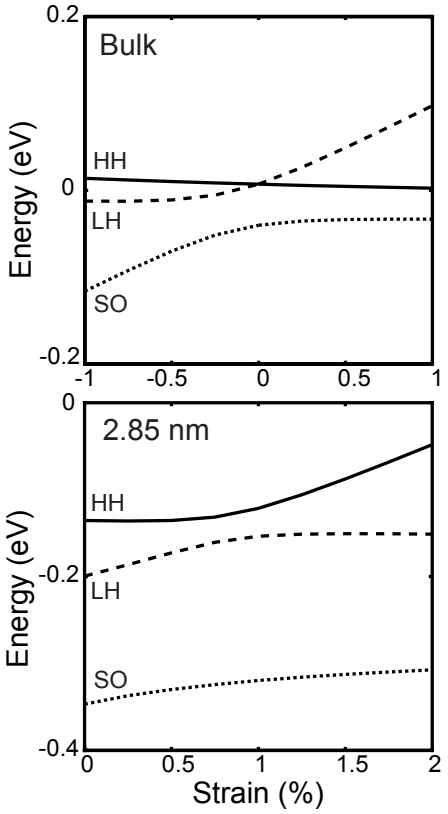


Figure 3-17: Effect of biaxial strain on the valence band energies in bulk silicon (top) and ultrathin body SOI (bottom). In UTB SOI the degeneracy of the valence band is already removed, due to the strong confinement. Unlike in bulk silicon, where the band splitting between the light and heavy holes increases monotonically, in UTB SOI and for tensile strains less than about 1%, these bands merge and then start to diverge for higher values of strain.

start to diverge again. This turnover strain level is pushed to higher values as the silicon layer is thinned, reaching values as high as 1% at the 3 nm. The net result is that tensile strain is not as effective in UTB SOI as in bulk silicon. The situation is qualitatively similar to the loss of the hole mobility enhancement in bulk silicon at high electric fields, where the band splitting caused by the strain is partially canceled by the confinement [155].

In bulk silicon, applying a biaxial tensile strain also modifies the shape of the band structure in a manner that reduces the effective mass of holes in the [110] channel direction. However, the dispersion relations are quite different in UTB SOI [156]. The mass anisotropy in unstrained material is much stronger in UTB SOI compared to bulk silicon due to strong confinement. A moderate tensile strain reduces the anisotropy, but is not enough to reduce the effective mass in the [110] channel direction significantly. Consequently, strain values in excess of 1% are needed in UTB SOI to reduce the effective mass of holes in the [110] channel direction. However, contrary to bulk silicon, where the effective mass in the [100] direction remains essentially constant, in UTB SOI it increases drastically once the effective mass in the [110] direction starts to decrease. The confinement-induced change in the effective mass observed here is qualitatively similar to what has been reported for hole inversion layer in bulk Si and under high electric fields [157].

3.5.5 Biaxial Compressive Strain

As can be seen from Figure 3-17, a biaxial compressive strain also lifts the degeneracy between the LH and HH bands, but favors heavy holes. Although the band splitting is small it can increase the mobility slightly by suppressing the optical phonon scattering. Preferential population of the HH bands is not a concern as most of the holes are in HH under quantum confinement due to higher quantization and DOS mass of this band. In fact, slight hole mobility enhancement is seen in early silicon-on-sapphire MOSFETs due to the presence of biaxial compressive strain [158]. However, since this strain degrades electron mobility by populating the Δ_4 valleys, provisions were sought later to relax the strain.

Figure 3-17 suggests that biaxial compressive strain can enhance hole mobility considerably in ultrathin SOI. Although, it is not clear how a biaxial compressive strain can be induced to a silicon layer, the beauty of this approach is that the band splitting is in the same direction that the quantum confinement is pushing the bands.

Figure 3-18 illustrates a more interesting situation. The band splitting in germanium under biaxial compressive strain is a lot more than the corresponding split in silicon. No wonder exceptionally high hole mobility has been observed in Ge channels grown on SiGe buffer layers (with smaller lattice constant) [39]. Similar to Si, the top energy band is HH, in harmony with the band splitting under quantum confinement, but with small and nearly symmetric effective mass near the top of the valence band as seen in Figure 3-19. However, for germanium layers less than about 10 nm, there is already enough band splitting to suppress the interband phonon scattering and addition of the biaxial strain is unlikely to add any benefit. In fact in ultrathin layers, it might decrease the mobility by excessive increase in the effective mass in the [110] direction. Hence, it would be a wise idea to use [100] as the channel direction. Alternatively, biaxially strained germanium layers can be used with additive unixial compressive strain in [110] direction to remove the wing of the HH $E-k$ diagram in this direction or as a starting material to obtain uniaxially strained Ge by preferential relaxation of the strain.

3.5.6 Germanium

Germanium has the highest bulk hole mobility among semiconductors. This is in part due to the smaller optical phonon energy and in part to smaller effective mass according to Table C.3. Hole mobility as high as 1000 $cm^2/V.s$ has been reported in Ge MOSFETs [136]. Recent publications, however, show maximum mobility of 250-300 $cm^2/V.s$ with oxynitride gate dielectrics [33,37] and 150 $cm^2/V.s$ with high-k dielectrics directly on Ge. Higher values have been reported, but with devices

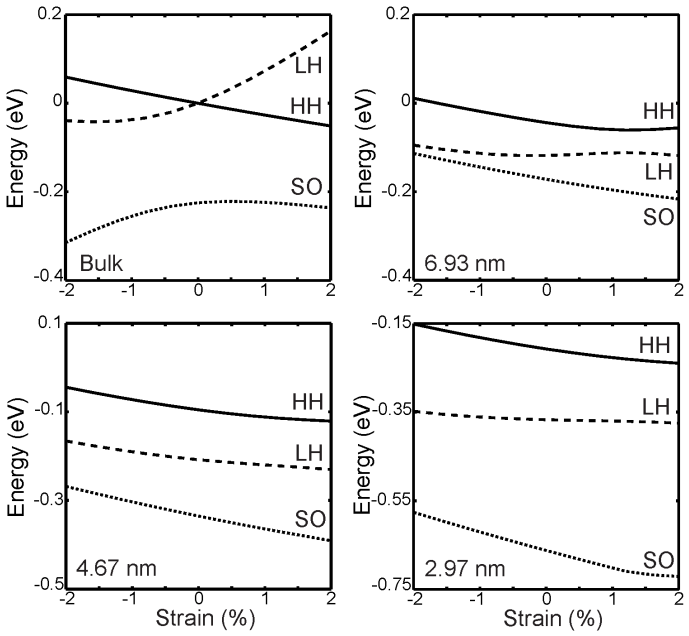


Figure 3-18: Effect of biaxial strain on the valence band energies in bulk germanium and ultrathin GOI. There is significant band splitting under biaxial compressive strain.

that have a silicon passivation layer [159][11] However, it appears that these levels of mobility enhancement compared to Si do not translate to a significant increase in the virtual source velocity. Short channel germanium MOSFETs with Si passivation and gate length down to 120 nm show a virtual-source velocity of around 4×10^6 cm/s, which is comparable to relaxed Si MOSFET at similar gate length and DIBL, despite 2.5× mobility enhancement compared to Si universal hole mobility [159]. Early Ge MOSFETs show more promising results. At 0.6 μm gate length for Ge PFETs reported in [137] the effective velocity is estimated to be around 4.5×10^6 cm/s. Virtual source velocity could be higher depending on the exact value of the source series resistance. Of course, this estimate involves some uncertainty and the devices in [137] seem to have relatively high DIBL. Yet, these results are somewhat encouraging and analysis of carrier transport in short channel Ge MOSFETs deserves

[11]Si passivation in [159] is achieved by low-temperature epitaxial growth of several monolayers of Si usually in ultrahigh vacuum. It has been reported that since this Si layer is under excessive stress, fluctuations in the Si thickness result in unacceptably high fluctuations in the threshold voltage. However, such technical difficulties might not be as sever with compressively strained Ge.

more attention.

Since the valence band in Ge is similar to that of Si, basic behavior of carrier transport enhancement with different mechanisms is essentially similar to that discussed in the above sections. Figure 3-19 shows the HH band structure in bulk Ge under different strain conditions or with different wafer orientations. Under biaxial compressive strain (1% here) the effective mass near the top of the valence band decreases significantly. Combined with the band splitting observed in Figure 3-18, this results in significant mobility enhancement. Interestingly, the band structure far from the top of the valence band is almost intact, not quite in favor of increasing the mobility. This might be an encouraging observation: Spectacular hole mobility enhancement observed in Ge under biaxial compressive strain could be mostly due to the decrease in the effective mass in the vicinity of the Γ point. While short channel transistors are yet to be fabricated on biaxially strained Ge to see whether mobility enhancement leads to a significant increase in the virtual source velocity, simulations can be performed to shed some light on the mechanisms involved. Alternatively, low temperature measurements can be used to see if hole mobility enhancement is observed at low temperatures where phonon scattering is not very important.

The mechanism for mobility enhancement under uniaxial compressive strain is similar to what is observed in Si: Reduction of the effective mass in the [110] channel direction. So, in principle it should lead to significant increase in the virtual source velocity. Figure 3-20 compares the enhancement in the ballistic velocity calculated in bulk Si and Ge subjected to uniaxial compressive strain and at a inversion charge density of 10^{13} cm^{-2}. Impressively high ballistic velocity is expected from Ge with uniaxial compressive strain. Similar calculations show that under biaxial compressive strain, Ge offers at most 2$\times$ velocity enhancement over relaxed silicon, which is comparable to what can be achieved with uniaxially strained silicon.

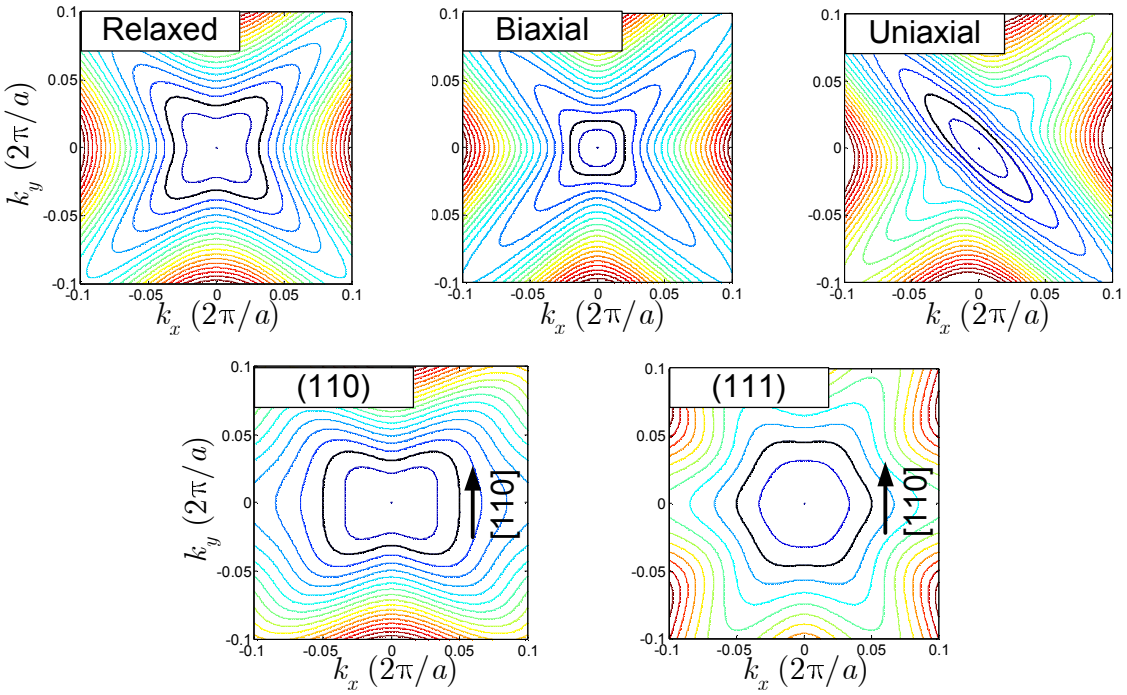


Figure 3-19: Heavy hole dispersion relation in (top) relaxed Ge, under 1% biaxial compressive strain, and under 1% uniaxial compressive strain with a (100) wafer orientation and (bottom) in relaxed Ge with (111) or (110) orientation.

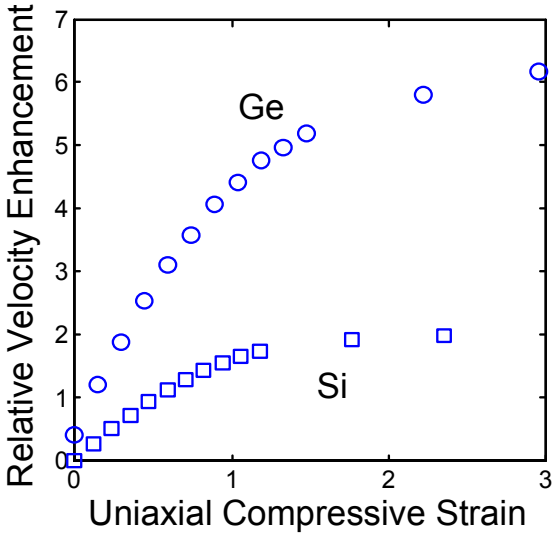


Figure 3-20: Ballistic velocity enhancement as a function of the uniaxial strain in silicon and germanium and at an inversion charge density of 10^{13} cm^{-2}. Band structures were calculated using $k.p$ method and the velocity was estimated using FETtoy [160].

3.6 Relationship Between Mobility and Velocity in Uniaxially Strained Si

Recent experimental data for deep sub-micron strain-engineered devices [12, 161] demonstrate that the saturation drain current, $I_{D\text{sat}}$, is more strongly correlated to the low field mobility, measured in the same short channel devices, than what was previously believed [105, 106]. To understand this dependence, it is more instructive to study the correlation between the virtual source velocity and mobility[12].

The relative change in the virtual source velocity can be written as the relative change in the ballistic velocity plus the change in the ballistic efficiency:

$$\frac{\partial v_{x0}}{v_{x0}} = \frac{\partial v_\theta}{v_\theta} + \frac{\partial B}{B} \tag{3.19}$$

With the power-law dependence between the ballistic velocity and mobility defined in Section 3.2.2, and taking into account the change in the critical length of backscattering discussed in Section 3.2.3, we have:

$$\frac{\partial v_{x0}}{v_{x0}} = [\alpha + (1 - \alpha + \beta)(1 - B)]\frac{\partial \mu}{\mu}. \tag{3.20}$$

Figure 3-21 shows the correlation between virtual source velocity, calculated based on the reported $I_{D\text{sat}}$, and mobility measured in the same short channel devices as reported in [12]. The ratio of the change in the velocity to that of mobility is much higher than 0.5. A similar observation is made for other strained devices (Figure 3-22) [86, 128, 162–165]. Here, change in the mobility is deduced from the change in the slope of the $R_{tot} - L_G$ characteristics and might contain some uncertainty. Caution should be taken drawing general conclusions. None of the methods used to apply mechanical stress to the channel and thereby improve the mobility and drive current produce a purely uniaxial stress along the channel [166]. The actual stress pattern

[12]The notion of mobility in such short-channel transistors, where the mean free path is comparable to the channel length, is a subject of controversy. However, a phenomenological mobility can always be extracted at low V_{DS}. Eq. (3.10) relates this mobility to the mean free path of carriers in the channel by using the scattering MOSFET model to match the low-V_{DS} drift-diffusion equation [104].

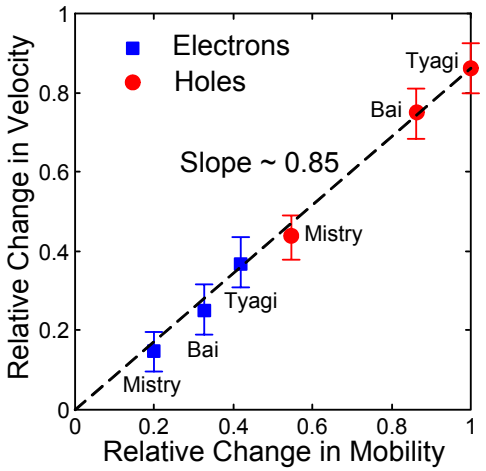


Figure 3-21: The relative change in the virtual source velocity vs. the relative change in the mobility based on the data given in [12]. The correlation ratio is much higher than the commonly accepted value of 0.5 [105, 106].

also depends on the device structure and geometry. Furthermore, when combined together, the effects of different methods do not add up linearly.

For PMOS transistors, the vertical component of the stress, if compressive, has negligible effect on the mobility and velocity, since the band splitting caused by this component is canceled out by the quantization. The splitting of the valence band due to a tensile stress in the vertical direction, although usually small, preferably populates the heavy holes [156]. So, it will result in lower ballistic velocity, while having small but positive effect on the mobility. For an NMOS transistor, the vertical stress component has significant impact on both mobility and velocity. Its effect can be viewed as a biaxial stress parallel to the wafer's surface, leading to more change in mobility than velocity. To gain the most performance possible from strain, one would like to engineer it in a way that translates to the highest possible velocity. In that sense, the strain distribution that reduces the transport effective mass is desirable.

An important question is whether the strong dependence of virtual source velocity and mobility as discussed above continues to hold at higher strain levels. This question has major consequences of the future of strain engineering: Should we continue increasing the strain level or should we seek alternative channel materials?

Based on experimental data, although hole mobility enhancement as high as 4×

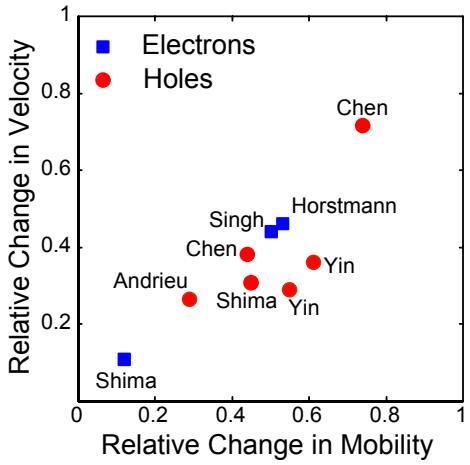


Figure 3-22: The relative change in the virtual source velocity vs. the relative change in the mobility for strain-engineered devices based on indirect data from [86,128,162–164] and mobility data from [165].

in uniaxially strained silicon [30, 89] and up to 8× with uniaxially strained (110) silicon [30] has been reported, the enhancement in the virtual-source velocity seems to saturate at 2-2.5×. Saturation of the velocity-mobility dependence in uniaxially strained PFETs can be understood based on the ballistic velocity-mobility dependence shown in Figure 3-15 and the band structure calculations in Figures 3-14. At higher strain levels the band structure near the top of the valence band does not change anymore. At a given hole density, the dispersion relation only up to the point where the band is occupied by the holes is important in determining the ballistic velocity. However, mobility depends to the shape of the band structure up to energies $\hbar\omega_{op}$ higher, where $\hbar\omega_{op}$ is the energy of the optical phonons. So, while mobility continues to increase at higher strain levels, velocity saturates once the band structure up to the Fermi wavenumber stops changing. Saturation of the velocity-mobility dependence is also seen in Figure 3-23 which collects virtually all experimental data available on strain engineered PMOS transistors.

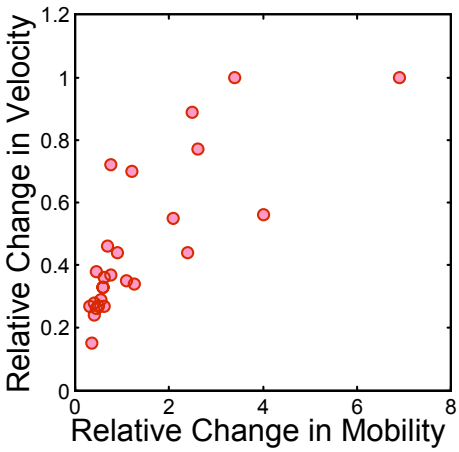


Figure 3-23: The relative change in the virtual source velocity vs. the relative change in the mobility for PMOS transistors. Generally the $\partial v_{xo}/v_{xo}/\partial\mu/\mu$ drops at higher levels of mobility enhancement and it seems that enhancement in virtual source velocity is limited to about 100%. This can be justified by the fact that at higher levels of uniaxial compressive strain only the band structure far from the Γ point changes. Velocity only depends on the shape of the band structure up to k_F, while mobility probes energies up to $\hbar\omega_{op}$ higher.

3.7 Conclusions

This section provided the guidelines for choosing a channel material, wafer orientation, and strain configuration to enhance the transistor performance. The significance of maximizing the quantization effective mass, while choosing proper in-plane masses was discussed based on different parameters that contribute to the performance.

The main conclusions for enhanced electron transport are:

- Biaxial tensile strain in silicon does not offer significant velocity enhancement since the $\approx 2\times$ mobility enhancement is mainly due to reduction in scattering rate. Biaxial strain can be preferentially relaxed to achieve uniaxial strain without the need to nitride stressors and the associated increase in the parasitic capacitance.

- Uniaxial tensile strain in silicon seems very promising as long as higher strain levels can be exerted on the device. Since theoretical band structure calculations do not agree with each other and with experimental data, it is not yet clear if modulation of the effective mass will continue at higher strain levels. Another

question is whether the band structure at higher energy levels is also modulated. Preliminary Monte Carlo simulations show that ultimately the velocity enhancement is saturated at about 2% of uniaxial strain since the velocity near the drain does not increase as much as the virtual source velocity does with increasing the strain.

- The performance offered by III-V semiconductors is limited mainly because of their small quantization effective mass.

- To avoid interband scattering and mobility degradation in germanium-on-insulator structures, (111) wafer orientation should be used. This orientation also offers the highest ballistic velocity in Ge. Nevertheless, in (111) bulk germanium, the ballistic velocity is only about 80% higher than relaxed silicon. This is significant when compared to state-of-the-art uniaxially strained silicon, especially if theoretically high mobility of electrons is demonstrated in Ge MOSFETs.

While the main conclusions for enhanced hole transport are:

- With uniaxially strained Si, the ballistic velocity enhancement is limited to about 2×, despite the fact that mobility enhancement of about 4× has been demonstrated. Hence, further increase of the strain level does not seem to provide major increase in the device performance.

- Without any strain, germanium only marginally improves hole velocity despite the fact that mobility is significantly higher than silicon. Biaxial compressive strain, although relatively simple to apply, offers only 2× velocity enhancement over relaxed silicon. Only with uniaxial compressive strain germanium is able to provide significantly higher velocities compared to the state-of-the-art silicon MOSFETs.

Despite the attractiveness of germanium as an alternative channel material, experimental data so far have been disappointing. Next chapter will examine the carrier mobility measured in germanium MOSFETs to determine what phenomena are limiting the mobility and what can be done to improve the device characteristics.

Chapter 4

Characterization of Electron Transport in Germanium Channel MOSFETs

As discussed in the previous chapter, germanium is of prime interest as an alternative channel material to offer superior transport properties compared to what can be achieved with silicon. However, despite considerable effort by several groups, the reported electrical characteristics of Ge channel MOSFETs have been unsatisfactory. In particular, germanium NFETs have generally exhibited very poor mobility, whereas mobility in PMOS transistors, although much higher, is still below what is expected theoretically.

The degradation of carrier mobility in Ge MOSFETs is partly due to the fact that the high-κ dielectric processes adopted for germanium are not yet optimal. On one hand, carrier scattering is enhanced at the high-κ interface possibly due to increased surface roughness scattering, remote phonon scattering [167], and Coulomb scattering [168] in the presence of charged traps at the interface or inside the gate dielectric. Moreover, charge trapping results in lower carrier density in the channel as compared with $C-V$ characteristics, making the conventional effective mobility measurement ambiguous. The split-CV technique, commonly used to determine the inversion charge, is unable to distinguish between the trapped and mobile charges.

Hence, the method usually underestimates the actual mobility in the channel. It is thus highly desirable to measure the intrinsic transport properties of the channel in order to understand what is limiting the mobility. This chapter reports some of the methods used in this work to characterize Ge channel transistors without being affected by the charge trapping as much as possible. Methods presented in the chapter are applicable to other exotic channel materials, such as III-V semiconductors, with some practical adjustments of the measurement details.

4.1 Impact of Charge Trapping on the Mobility Extraction

The effective mobility in the MOSFET channel is determined by measuring the drain current in the linear region:

$$\mu_{\text{eff}} = \frac{L}{W} \cdot \frac{I_D(V_G)}{V_D Q_{\text{inv}}(V_G)} \tag{4.1}$$

where the inversion charge, Q_{inv}, is usually obtained by integrating the split-CV characteristics:

$$Q_{\text{inv}}(V_G) = \int^{V_G} C_{gc}(V_G) dV_G, \tag{4.2}$$

and C_{gc} is the gate-channel capacitance.

There are two sources of error when determining the inversion charge: First, the interface traps might respond to the AC signal used for the capacitance measurement and result in an additional capacitance component, C_{it}, effectively parallel to C_{ox}. This can be observed as a frequency dispersion in the $C-V$ characteristics and the error can be minimized by using higher frequencies such that the interface traps cannot respond to the AC signal. Second, the traps can follow the DC voltage, so that a change in the gate voltage results in less change in the inversion charge than would be expected theoretically and in the absence of the traps. This effect is seen as a stretch-out of the $C-V$ curve and results in overestimation of the inversion charge.

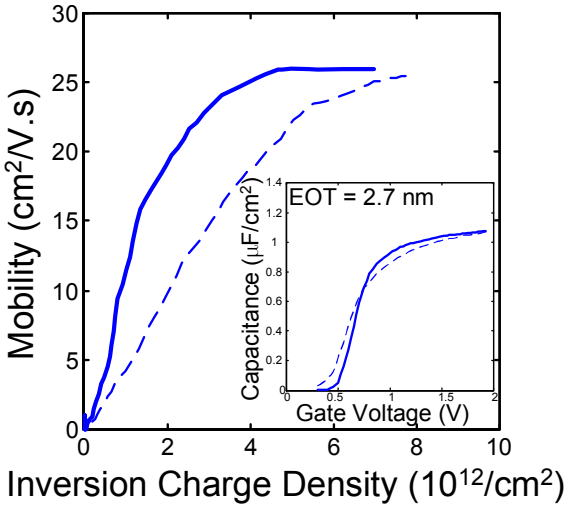


Figure 4-1: Correction of the mobility extracted from DC $I-V$ data and split-CV measurements according to [169]. Dashed line shows the mobility extracted without the correction whereas solid line shows the corrected values. The inset shows the measured (dashed line) and simulated (solid line) $C-V$ characteristics.

Zhu *et al.* [169] proposed a correction scheme to estimate the "true" inversion charge from the measured split-CV characteristics by using the ideal $C-V$ curve. This method assumes that the inversion capacitance at a given inversion charge density (or equivalently at a given surface potential) is unchanged whether interface traps are present or not (as long as they do not respond to the AC signal). Note that the threshold voltage shift due to fixed charges does not affect this method, which only relies on the shape of the $Q_{\mathrm{inv}}(C_{gc})$ curve. Although this method seems straightforward, calculating the ideal $C-V$ characteristics poses new problems. To obtain the actual shape of the $C-V$ curve requires self-consistent simulations with accurate band-structure parameters, which is not trivial for new channel materials and especially heterostructures.

Figure 4-1 shows an example of applying this method to Ge NMOS transistors. Schroedinger-Poisson simulations where performed with parameters adjusted for (100) Ge to obtain the ideal $C-V$ and thereby $Q_{\mathrm{inv}}(C_{gc})$ characteristics. As can be seen, the correction is significant at lower gate voltages but the difference diminishes at higher gate voltages.

Alternatively, one could measure the density of trapped charges at each gate voltage and subtract from the total measured charge to obtain the true inversion charge. However, charge trapping and detrapping happen during the measurements and the density of trapped charges depend on the transient bias conditions; so, it is important to use consistent conditions for all measurements.

It should be noted that the above method does not measure the intrinsic channel mobility in the absence of the charge trapping. Rather, it only corrects the inversion charge estimation for the presence of the slow traps. In the following sections methods are presented that extract near intrinsic transport properties of the channel and show how mobility is affected by charge trapping.

4.2 Pulsed $I-V$ Measurements

As mentioned above, charge trapping and detrapping precludes correct measurement of the device characteristics. The intrinsic $I-V$ characteristics can be obtained if the voltage sweep is performed fast enough so that the traps cannot respond to it [170]. One could either apply a train of pulses with increasing height to the gate and measure the drain current at the top of each pulse, or perform a "single" pulse measurement, in which pulses are applied to the gate and the $I-V$ data are collected during the ramp times [170]. Near intrinsic characteristics can be obtained if the rise and fall times of the pulse are short enough and the device does not spend too much time in inversion. This method also offers the possibility of determining the time-scale in which the traps respond to the gate voltage by varying the pulse width and rise and fall times.

Figure 4-2 shows some of the schemes proposed for pulsed $I-V$ measurements. In the most common scheme shown in Figure 4-2 (a) the potential drop across a resistor connected between the drain terminal and a supply voltage is measured to calculate the drain current. To minimize the signal reflection at the circuit discontinuities, 50 Ω-terminated probes are used. In our implementation, instead of connecting a load resistance between the drain and a DC supply voltage, we used the output resistance

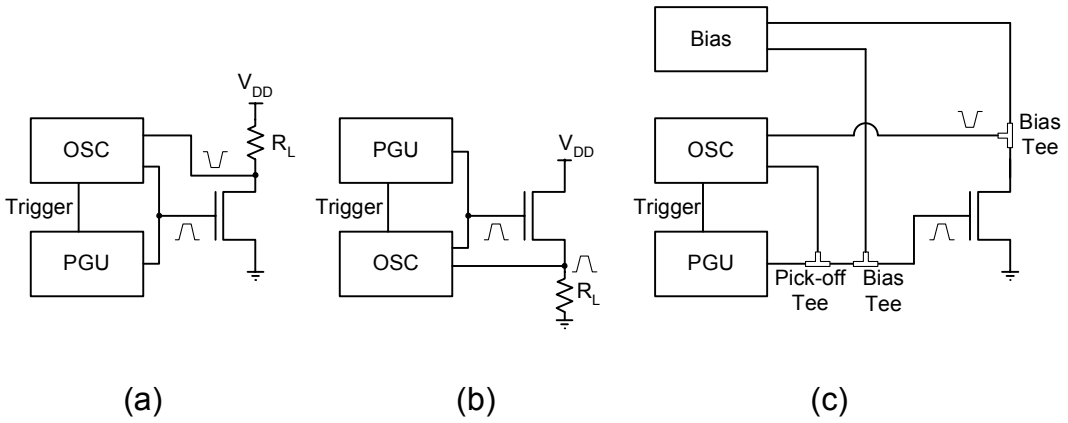


Figure 4-2: Different configurations used for pulsed I-V measurements. (a) An inverter configuration where the voltage drop across the load resistance connected to the drain is measured to give the drain current. (b) A source follower configuration where the source current is measured across a load resistance and has higher bandwidth compared to (a). (c) A high-frequency configuration that uses bias tees to decouple DC and AC signals.

of the pulse generator (50 Ω) as the load and simply kept the generator output at a constant level (V_{DD}). This simplifies the bias circuitry, which otherwise requires proper filtering of high-frequency signals at the supply voltage terminal.

An alternative approach is to measure the source current by placing a resistor on the source terminal, as shown in Figure 4-2 (b). This method has potentially higher bandwidth as it has a source follower configuration and is free of the Miller effect. However, the data should be corrected for the current dependence of V_{GS}.

The third scheme shown in Figure 4-2 (c) [170], is essentially a high-frequency cabling using bias tees to decouple the signal from the bias circuitry. Frequencies from 1 MHz to 10 GHz pass through the bias tee and lower frequencies turn to the bias terminal. Although this scheme has very high bandwidth and is immune to the noise from the bias circuit, it is not suited for studying the contribution of slow traps as only transient behavior with a time-scale smaller than about 1 μs can be observed.

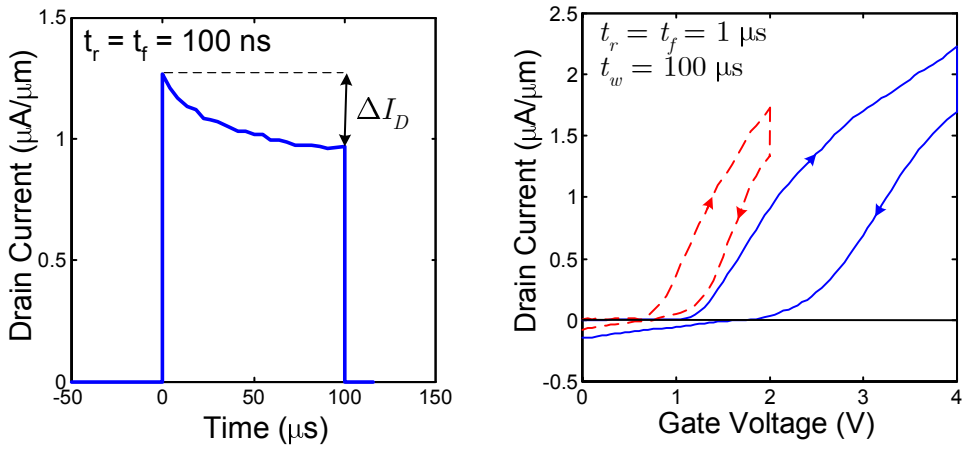
Figure 4-3 shows sample pulsed $I-V$ data collected from Ge NMOSFETs using the configuration shown in Figure 4-2(a), featuring several non-ideal characteristics. Pulses with varying rise and fall times of $t_r = t_f = 100$ ns to $t_r = t_f = 1$ μs, pulse

width of $t_w = 100$ μs, and period long enough to ensure complete detrapping ($T = 1$ ms) were applied to the gate and the drain current was measured. The transient characteristics are shown in Figure 4-3 (a), whereas the data during rise and fall ramps were converted to the $I - V$ characteristics in (b). The data in (a) show a gradual drop in the drain current with a time constant of about 50 μs, due to the presence of slow traps. This can be seen in (b) as a shift to the right for the ramp down trace, which can be viewed as a shift in the threshold voltage which depends on the pulse amplitude as seen in Figure 4-3 (b).

In addition to the threshold voltage shift, the $I - V$ characteristics in (b) exhibit some distortion as the pulse amplitude is increased. The traces recorded at different pulse magnitude are neither parallel, nor is the ramp down trace parallel to the ramp up trace. The dependence of the $I - V$ characteristics on the pulse amplitude could either be due to generation of some defects in the oxide or at the oxide-semiconductor interface upon electrical stressing, or be a symptom of fast traps that respond to the gate signal with a time scale comparable to the ramp times. The distortion of the ramp down trace can be attributed to the presence of slow traps that are gradually charged when the transistor is in the inversion regime. Although the shift in the threshold voltage can be used to quantify the density of trapped charges in each case, we prefer to directly measure the charge trapping as will be discussed in Section 4.4.

4.3 Mobility Extraction from Pulsed $I - V$ Data

Pulsed $I - V$ measurements were used to extract near intrinsic electron mobility in n-channel germanium MOSFETs. Details of the device fabrication has been reported in [171]. Long-channel ring transistors were fabricated on (100) bulk germanium. Some of the wafers were implanted with phosphorus with doses in the range of $1 - 4 \times 10^{12}$ cm^{-2} through an 11 nm CVD SiO_2 screen oxide and annealed at 500°C for 60 s in nitrogen [171]. The screen oxide was then removed and the wafers were cleaned in 6:1 DI:HCl for 5 min. and rinsed. Wafers that did not receive the phosphorus implant were cleaned in H_2O_2:NH_4OH:DI 1:1:5 for 15 s, followed by HCl dip and



(a) (b)

Figure 4-3: Pulsed $I-V$ measurements results for a typical Ge NMOSFET. (a) A trace of the drain current recorded with a pulse amplitude of 1.5 V, $t_r = t_f = 100$ ns, and $t_w = 100$ μs. The drain voltage drops gradually as a result of slow charge trapping. (b) The data recorded during the ramp up and ramp down converted to the $I_d - V_g$ characteristics. In this case $t_r = t_f = 1$ μs. Dashed lines and solid lines correspond to measurements performed with a pulse amplitude of 2 V and 4 V, respectively.

rinse. An *in situ* ALD WN/Al_2O_3/AlN gate stack was then deposited and patterned with a dry etch. Source/drain implantation was performed (P, 1×10^{15} cm^{-2}, 25 keV), followed by PECVD oxide deposition, S/D activation, and Ti/Al contact deposition and patterning [171]. Conventional DC $I-V$ and split-CV measurements show enhanced electron mobility as the phosphorus passivation dose is increased [171]. In this section the mobility extracted from pulsed measurements are compared to better understand the carrier transport in these devices.

Pulsed $I-V$ measurements were performed on the smallest transistors on the wafer ($W/L = 180/5$ μm), with a load resistor $R_L = 25\Omega$ connected to the drain. A train of trapezoidal pulses with increasing amplitude were applied to the gate and the voltage at the drain terminal was measured and converted to I_d. Gate leakage was negligible but measurements were performed with $V_{DD} = \pm100$ mV connected through R_L to correct for any possible leakage.

Figure 4-4 compares the DC and pulsed $I-V$ characteristics measured on transistors that received different doses of phosphorus implant prior to high-κ deposition.

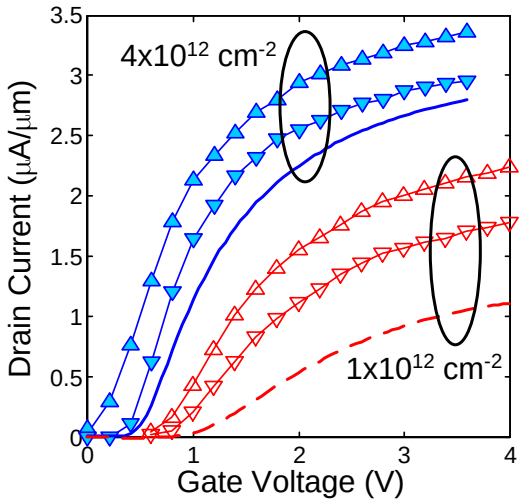


Figure 4-4: Comparison of $I_d - V_g$ characteristics from DC (lines) and pulsed (symbols) measurements for devices that received different doses of phosphorus implantation prior to high-κ deposition. Measurements were performed on ring transistors with $W/L = 180/5$ μm. Pulsed measurements were done by applying a train of pulses with increasing pulse height, $t_r = t_f =$ 100 ns, and $t_w = 100$ μs. $\triangle$ represents the measurements at the beginning of the pulse, while $\triangledown$ indicates the values after t_w.

I_d measurements were performed at the beginning and end of pulses with a width of 100 μs and rise and fall time of 100 ns. Devices that received the lowest dose have considerably lower current and the ratio between the pulsed and DC measurements is about 2. With an implant dose of 4×10^{12} cm^{-2}, DC and pulsed measurements show an improvement of more than 2$\times$ and 1.5$\times$, respectively and the ratio between the pulsed and DC measurements drops to about 1.2. Moreover, it appears that at lower gate voltages the difference between the DC and pulsed characteristics is merely a shift in the threshold voltage. In other words, although the density of trapped charges is approximately 1.2×10^{12} cm^{-2} at low gate voltages, as estimated from the shift in the threshold voltage, carrier mobility is not affected by Coulomb scattering of these charges. At higher gate voltages, degradation of the $I - V$ characteristics occurs, similar to the behavior in the sample implanted with lower doses. This suggests that multiple mechanisms of charge trapping are present and phosphorus implantation only mitigates mechanisms responsible for charge trapping at lower gate voltages.

Since split-CV measurements are also contaminated with charge traps and hence

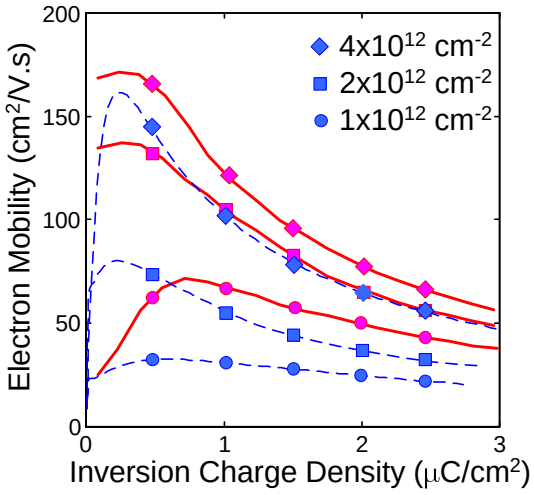


Figure 4-5: Mobility extracted from DC (dashed lines) and pulsed (solid lines) measurements and on devices that received different doses of phosphorus implant. As the implant dose increases both DC and pulsed mobility values increase, but the DC values approach the pulsed measurement results.

show some degree of skew in weak inversion, to extract the mobility from pulsed $I-V$ measurements the ideal inversion $C-V$ curves from self-consistent simulations were used. Care was taken to match the threshold voltage in the simulated $C-V$ curve with that of measured pulsed $I-V$ characteristics. Alternatively, the inversion charge can be directly estimated as will be discussed in the next section. Split-CV data were used to extract the mobility from DC measurements, since the $I-V$ data are equally skewed due to charge trapping and gradual increase of the threshold voltage. The DC ramp for the $I-V$ and $C-V$ measurements were matched to minimize the error due to difference in the amount of charge trapping.

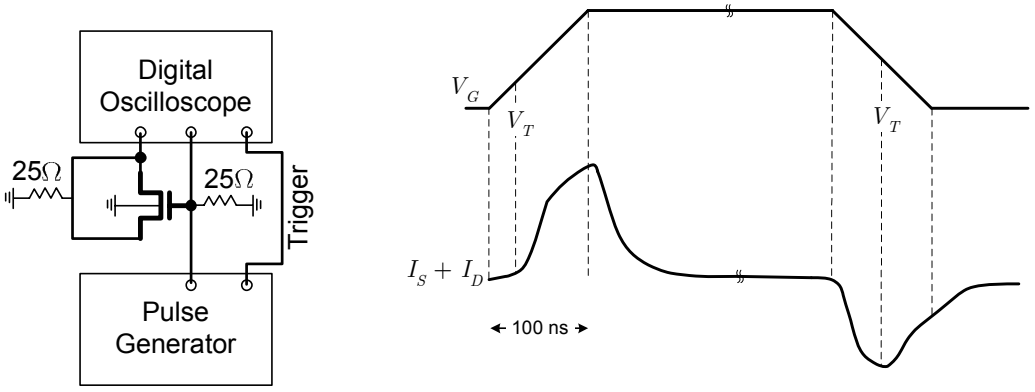
Figure 4-5 shows the extracted mobility from the DC and pulse measurements for samples with different doses of phosphorus implant. Both DC and pulse measurements show that mobility increases by increasing the implant dose. However, the difference between mobility curves extracted from DC and pulsed $I-V$ measurements almost vanishes as the phosphorus dose is increased.

4.4 Direct Measurement of the Inversion Charge

In the previous section, simulated $C-V$ characteristics were used to extract mobility from pulsed $I-V$ data. That approach involves some degree of uncertainty on how to determine the equivalent oxide thickness and how to match the threshold voltage of the $C-V$ and $I-V$ data. The inversion charge can be measured directly by integrating the transient current that follows into or out of the source and drain terminals during the V_g ramps. Figure 4-6 (a) shows a schematic diagram of the measurement setup. A train of pulses are applied to the gate terminal and the source and drain are connected together and to a digital oscilloscope; the current through these terminals is converted to a voltage drop across the parallel combination of the oscilloscope input resistance (50 Ω) and the coaxial probes (50 Ω per each). The inversion charge at the ramp-up or ramp-down is determined by digital integration of the current during that time period. The exact transient behavior of the source/drain current, which depends on the channel resistance that varies during the ramp time, is not important as only the area under the signal is needed. Singh *et al.*, proposed a similar approach to directly measure the inversion charge [172]. Instead of digital integration of the S/D current, they used a load capacitor to integrate the current.

Figure 4-6 (b) shows some sample traces obtained during ramp up and ramp down on a Ge NMOS transistor. It is interesting to note that the area under the ramp-up and ramp-down bumps are not equal. In other words, charge transfered to the channel to form the inversion layer at the beginning of the pulse is higher than the charge returned to the source/drain at the end of the pulse. The difference can be attributed to the presence of the slow traps. The trapped charges detrap at a later time and either recombine with the majority carriers in the substrate or diffuse to the source/drain.

Figure 4-7 shows sample $Q-V$ characteristics obtained on Ge NMOSFETs by directly measuring the inversion charge as described above. The charges at the beginning and at the end of a pulse with $t_w = 100$ μs are not equal as reflected in Figure 4-7 (a); there is a positive shift in the threshold voltage and a drop in the ca-



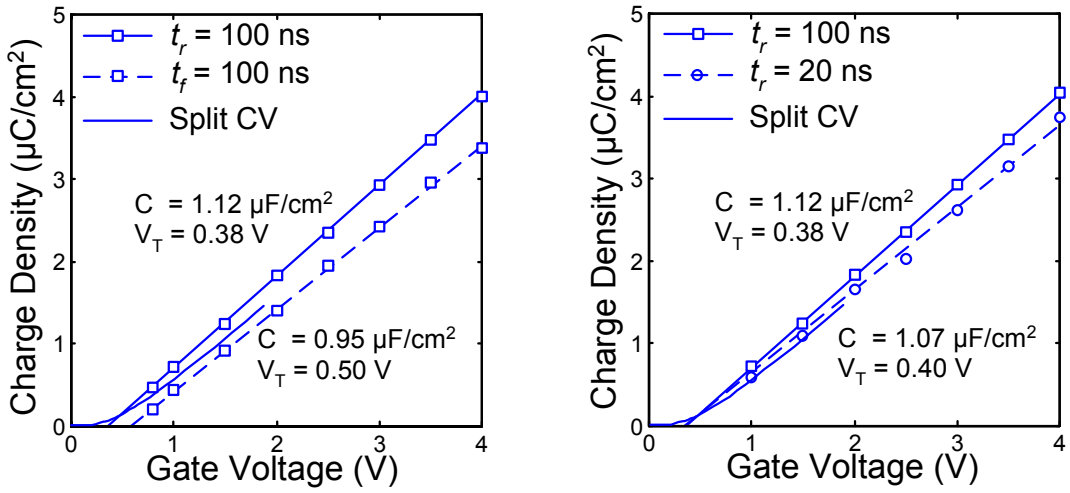
(a) (b)

Figure 4-6: (a) A schematic diagram of the setup used to directly measure the inversion charge by integrating the transient source/drain current during ramp up and ramp down. (b) A sample trace of the data collected during ramp times on a Ge NMOSFET with $t_r = t_f = 100$ ns, $t_w = 100$ μs, and pulse amplitude of 4 V. The approximate values of the threshold voltage extracted from the corresponding $I - V$ characteristics are denoted. The presence of a series resistance, composed of the distributed channel resistance and the S/D series resistance, distorts the shape of the signal. However, as fas as the total inversion charge at a given voltage is concerned, only the area under the curve is important.

pacitance between the rise an fall time of the pulse. In (b), plotted are measurements performed at the beginning of the pulse, but with different rise times of $t_r = 20$ ns and 100 ns. A small drop in the inversion capacitance is observed in this case that can be attributed to the presence of very fast traps. This canbe explained based on the data presented in Figure 4-8 that shows sample traces of the source/drain current during switching with very small rise and fall times. Two distinct bumps are observed in this plots: one with a time scale comparable to the pulse rise time, of course somewhat lagging due to the presence of the series resistance[1] and one with a time scale larger than about 100 ns. It is speculated that the first bump is due to the actual inversion charges, whereas the second one represents the trapped charges that respond to the input signal but with a larger time constant. The difference in the slopes in Figure 4-7 (b) can be attributed to the presence of fast traps that can contaminate the $Q-V$ characteristics measured with a rise time of 100 ns, but are less deteriorating to the measurements performed with $t_r = 20$ ns. However, to precisely quantify the density of this type of fast traps, measurements should be performed on transistors with shorter gate length and reduced S/D series resistance to make sure the inversion charges respond fast enough to the gate signal, so that they can be distinguished from the traps.

The difference between the $Q - V$ characteristics recorded at the beginning and the end of the pulse can be plotted as the density of the trapped charges. The time scale of traps can be also examined by varying the pulse width. Figure 4-9 shows the trap density as a function of the gate voltage for samples that received phosphorus implant with different doses as passivation. Measurements are performed with a pulse width of 100 μs. So, both fast and slow traps are present in these plots. While the trap density is almost independent of the gate voltage at lower voltages, it increases linearly at higher voltages. As the gate voltage is increased the probability of the electrons to tunnel through the interfacial dielectric and communicate with the traps

[1]Assuming that the total series resistance for the devices under study is about 2000 Ω, the time constant for the inversion charges to respond to the gate signal is about 20 ns. Due to the non-linear dependence of the channel resistance it is difficult to decouple the effect of the R-C delay in the current traces with very small rise and fall time.



(a) (b)

Figure 4-7: The inversion charge measured directly by integrating the S/D current on a Ge NMOSFET. (a) at the beginning and end of a 100 μs pulse. (b) At the beginning of the pulse with different rise times.

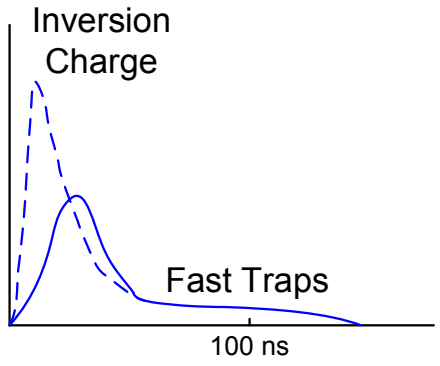


Figure 4-8: A sample trace of the inversion current recorded during the pulse rise time with $t_r = 10$ ns and 20 ns.

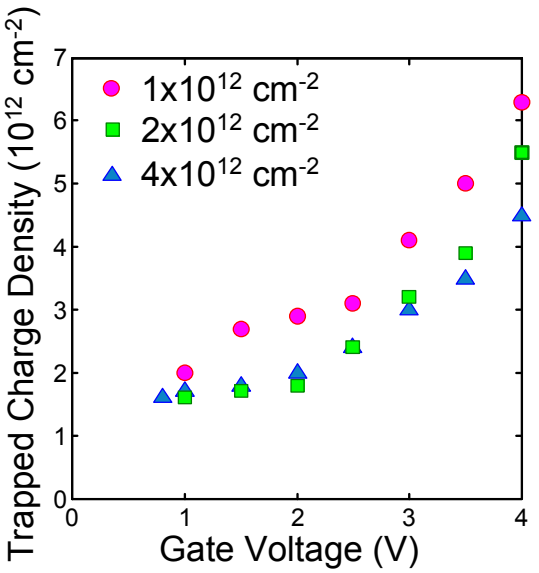


Figure 4-9: The density of trapped charges estimated by subtracting the inversion charge at the beginning and end of the pulses with $t_w = 100$ μs and for Ge NMOSFETs that received different doses of phosphorus implant as a passivation prior to high-κ deposition.

inside the dielectric increases and hence there is an increase in the density of trapped charges. This method is further used in Section 4.6 to study the charge trapping in germanium MOSFETs and its impact on electron mobility.

4.5 Charge Pumping

Since charge pumping is widely used to characterize interface state densities in MOSFETs [173], it was employed to study Ge MOSFETs in this work. As shown in Figure 4-10, a train of pulses is applied to the gate and the DC current at the substrate terminal is measured, while the source and drain are grounded. When the transistor is in the inversion regime some of the minority carriers provided by the source and drain are trapped. Once the transistor is switched to the accumulation, some of these trapped charges have a chance to detrap and recombine with the majority carriers. This recombination current contributes to the substrate current, which is measured to give an estimate of the density of the trapped charges.

As illustrated in Figure 4-10, two measurement methods are possible. In the first

method, the pulse amplitude is kept constant while the base voltage, V_{base}, is swept. The charge-pumping current is then plotted as a function of the base voltage. The charge pumping current, I_{CP} is only detected for $V_{\text{base}} + V_A > V_{fb}$ and $V_{\text{base}} < V_T$. The maximum charge pumping current gives the average density of interface traps, $N_{\text{it}} = I_{\text{CP}}/qAf$, where q is the electron charge, A is the gate area, and f is the pulse frequency.

In the second approach [174], the base voltage is kept constant so that the transistor is in accumulation when $V_{GS} = V_{\text{base}}$ and the pulse amplitude is varied. In this case, the current is often plotted as a function of V_{top}. Since the device goes deeply in the inversion regime, depending on the frequency of the pulses used, some of the carriers might have enough time to communicate with the traps inside the high-κ dielectric. Hence, this method is believed to give the density of the oxide traps, $N_{\text{ot}} = I_{\text{CP}}/qAf$. If the oxide is free of defects, the signal should flatten at higher gate voltages and exhibit no frequency dependence. In the presence of oxide traps, the extracted density of trapped charges increases with the gate voltage, as carriers have more chance to tunnel through the interfacial oxide. Also, as the pulse frequency is reduced the carriers have a chance to communicate with traps deeper inside the oxide. So, this method is very attractive in determining the spatial distribution of the oxide traps [175, 176]. However, quantifying the depth-frequency relationship of the charge trapping requires careful modeling of the tunneling through the interfacial oxide and involves some uncertainty.

Figure 4-11 shows the charge-pumping measurement results for Ge NMOSFETs with different doses of phosphorus passivation implant. In both methods a small reduction in the trap density upon increasing the P dose is visible. Using the conventional charge pumping the peak trap density is only about 5×10^{11} cm^{-2} [Figure 4-11(a)], an order of magnitude less than the traps measured by the variable-amplitude method at high gate voltages [Figure 4-11(b)]. Therefore, the charge measured by the variable-amplitude method cannot be located at the germanium-dielectric interface. Furthermore, the trap density measured by the variable-amplitude increases significantly by increasing the pulse width, whereas the data from fixed-amplitude method

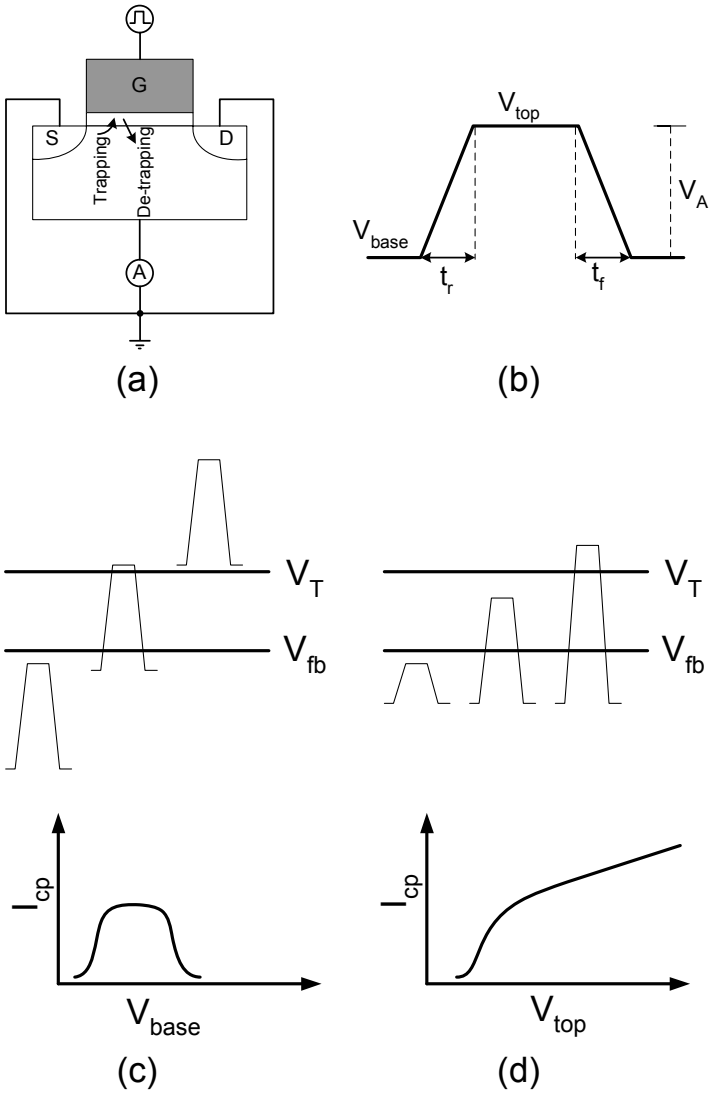
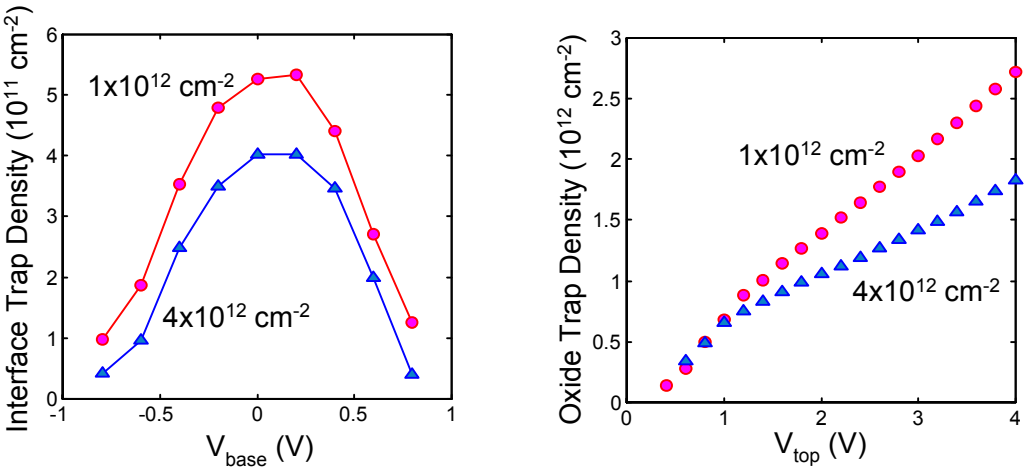


Figure 4-10: A schematic of charge pumping measurement on NMOSFET. (a) Measurement configuration where a train of pulses are applied to the gate to switch the transistor between accumulation and inversion and the DC current at the substrate terminal is measured with source and drain terminals grounded. When the transistor is biased in inversion, $V_G = V_{\text{top}}$, some of the electrons from the inversion layer are trapped in the interface and/or bulk high-κ states. When the transistor goes back to the accumulation, $V_G = V_{\text{base}}$, some of these electrons are de-trapped and recombined with holes in the substrate to create the charge pumping current, I_{cp}. (b) Parameters that define the shape of the charge pumping pulse. Two possibilities for the measurement are shown: (c) In the fixed-amplitude CP where the pulse amplitude is constant (and larger than $V_T - V_{fb}$) and V_{base} is varied. The interface trap density as a function of the band bending is extracted. (d) In the variable-amplitude CP, the base voltage is kept constant and the pulse amplitude is varied.



(a) (b)

Figure 4-11: Density of trapped charges extracted using (a) fixed-amplitude and (b) variable-amplitude charge pumping for Ge NMOSFETs with different doses of phosphorus passivation implant. For fair comparison, the data for higher phosphorus dose in (b) are shifted to adjust for the difference in the threshold voltage. The pulse amplitude was kept at 1 V in (a), while the base voltage was kept at -1 V in (b). The measurements were performed with a frequency of 1 MHz in (a) and 100 kHz in (b).

were essentially independent of the signal frequency (data are not shown in Figure 4-11 for clarity). It can thus be speculated that charge trapping is mostly due to the charging and discharging of the defects located inside the high-κ dielectric [174] and by tunneling through the interfacial dielectric. The tunneling probability increases with the gate voltage as seen in Figure 4-11(b).

Despite the attractiveness of the charge pumping method, its usefulness in studying charge trapping in Ge MOSFETs is limited. On one hand the background substrate current in samples studied in this work was high, usually 10-50 nA. This sets a lower limit on the frequency of the charge pumping pulses as the charge pumping current is proportional to frequency. On the other hand, due to the relatively high channel conductivity some of the charges that are detrapped when the transistor is in the accumulation regime diffuse to the S/D instead of being recombined and contribute to the substrate current. Hence, we decided to employ the method proposed in Section 4.4 in order to measure the density of the trapped charges.

4.6 Correlation Between Charge Trapping and Mobility Degradation

The method proposed in Section 4.4 to determine the density of trapped charges by subtracting the inversion charge at the beginning and end of the pulses provides a straightforward way to investigate the impact of charge trapping on the mobility. Two sets of the measurements are required. On the first pass, the inversion charge at the beginning and end of the pulses, at a given pulse amplitude with increasing pulse width is measured with the setup shown in Figure 4-2. The pulse period is kept large enough (5 ms) to ensure detrapping of all trapped charges. On the second pass, the drain current at the beginning and end of the pulses is measured with the same pulses with the setup shown in Figure 4-6.

Figure 4-12 shows the density of trapped charges as a function of the pulse width and at different gate voltages for a germanium NMOSFET. For smaller gate voltages, the density of trapped charges increases logarithmically with the pulse width. At higher gate voltages, two distinct regimes are visible, both following a logarithmic dependence on the pulse width. The first charge trapping mechanism, which is dominant for t_w up to about 20 μs, is a weaker function of the pulse width and somewhat increases with the gate voltage. The second mechanism, responsible for charge trapping at $t_w > 20$ μs, is a stronger function of the pulse width and increases significantly with the gate voltage.

It is speculated that the fast charge trapping corresponds to the traps located at the germanium-dielectric interface, which are thus in fast interaction with mobile charges. The slow traps are located inside the high-κ dielectric and thus the probability of the carriers to tunnel through the interfacial dielectric and get trapped increases with the gate voltage.

According to the theory of Heimann and Warfield [177], the distance of the filled trap centers from the interface is given by:

$$x_m = \frac{1}{2\kappa_0} \log(t_m \sigma_n \overline{v} n), \tag{4.3}$$

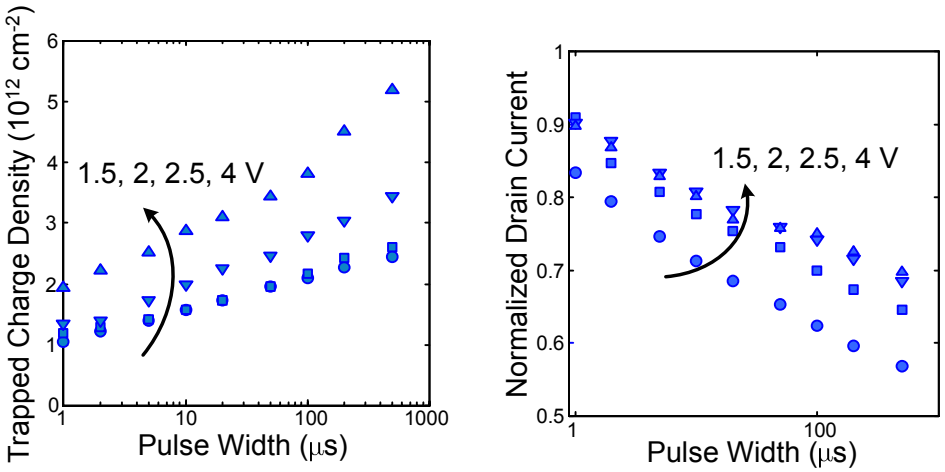
where $\kappa_0 = 2\lambda_n$ is the decay constant, with λ_n the attenuation coefficient, t_m is the tunneling time, equal to the pulse width in this case, σ_n is the capture cross section, $\overline{v}$ is the thermal velocity, and n is the electron concentration. The density of the traps at each depth, x_m, is obtained by differentiating the total trap density with respect to x_m:

$$N_t(x_m) = \frac{1}{\lambda_n \Delta E_t} \frac{dN_t}{d \log t_m}. \tag{4.4}$$

Here ΔE_t represents the energy gap in which traps will be filled by electrons. With a rise/time of 100 ns, as used in this work, $\Delta E_t = 0.7$ eV [178]. With parameters given in [176], the two regimes of fast and slow traps observed in Figure 4-12(a) and for the data at 4 V, correspond to a depth range of roughly 0.8 – 1.2 nm and 1.2 – 1.4 nm, respectively. The density of traps in the two regions is estimated to be roughly 1×10^{20} cm^{-3} and 2.4×10^{20} cm^{-3}, respectively. Of course, these numerical estimates should be taken with some caution as the parameters in [176] involve some uncertainty[2]. Higher or lower frequencies are required to obtain the depth profile of the traps closer to or farther away from the interface. In this work, the upper limit of the pulse frequency was limited by the device size ($L_G = 5\mu$m), while the lower bound was determined by the specifications of the digital oscilloscope.

Figure 4-12(b) shows the drain current at different gate voltages normalized to the current at the beginning of the pulse. The degradation of the current is more severe at lower gate voltages, where mobility depends more strongly on the Coulomb scattering. The $I_d - t$ data can be used in conjunction with the inversion charge density measured at each pulse width to plot the dependence of the electron mobility on the density of trapped charges, as shown in Figure 4-13. At lower gate voltages (as seen for the data for $V_g = 1.5$ V), electron mobility depends on the density of trapped charges for all time points. At higher gate voltages (such as the data for $V_g = 4$ V), mobility depends only on the density of the trapped charges during the first 10 μs

[2]In fact, the two depth regions could well correspond to the AlN and Al_2O_3 layers, respectively. Measurement on devices fabricated with different thickness of the AlN layer can be used to calibrate the model.



(a) (b)

Figure 4-12: Pulse-width dependence of the density of (a) trapped charges and (b) normalized drain current measured on Ge NMOSFETs and with different pulse amplitudes. For smaller gate voltages, the density of trapped charges increases logarithmically with the pulse width. For larger gate voltages, two distinct regimes are visible each exhibiting a logarithmic dependence on the pulse width, corresponding to the fast and slow traps. The drain current at each gate voltage is normalized to the current at the beginning of the pulse and shows stronger degradation at lower gate voltages.

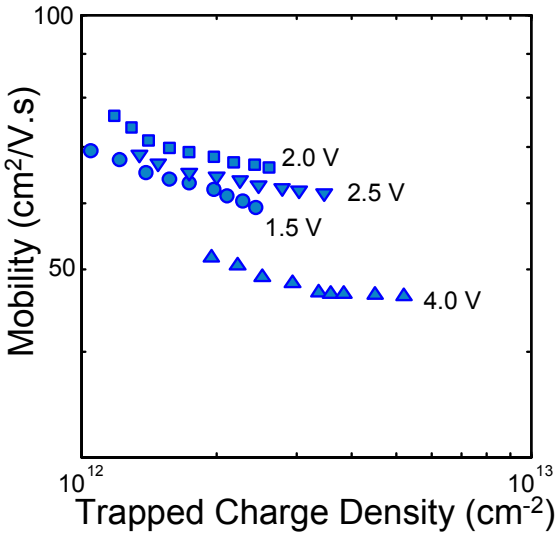


Figure 4-13: The dependence of the electron mobility on trapped charge density for a Ge NMOSFET and at different gate voltages. At lower gate voltages, electron mobility depends on the density of the trapped charges for all time points. At higher gate voltages, mobility depends only on the density of the trapped charges during the first 10 μs or so. Combined with the data in Figure 4-12 (a), this figure suggest that mobility only depends on the density of fast traps and fast traps are dominant at lower gate voltages.

or so[3]. From Figure 4-12(a) it was deduced that the charge trapping at lower gate voltages is dominated by fast traps, while at higher gate voltages a combination of fast traps (for time scales less than about 10 μs) and slow traps (at longer time scales) are responsible for charge trapping. Hence, the mobility data presented here suggest that electron mobility only depends on the density of fast traps. This is in agreement with the assumption that fast traps are located at the semiconductor-dielectric interface, close to the inversion centroid, and hence strongly scatter the carriers. Slow traps are located inside the high-κ gate dielectric, further away from the inversion layer, and hence do not affect carrier transport significantly.

To explore the dependence of the mobility on the density of fast traps in more detail, it is essential to measure the device characteristics at time scales of less than 1 μs.

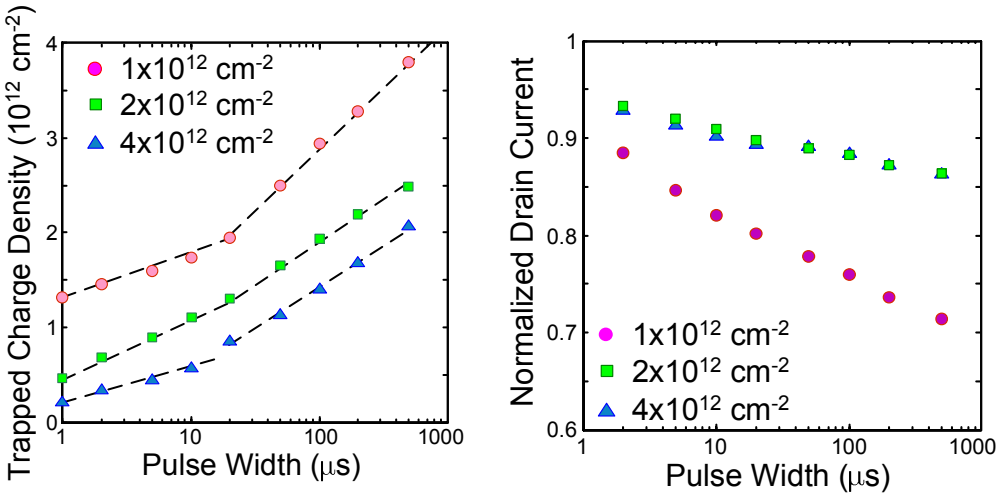
[3]Note that the non-monotonicity of the mobility data with respect to the gate voltage is simply due to the fact that mobility initially increases with the inversion charge density as the Coulomb scattering is screened and then drops as the phonon scattering and surface roughness scattering increase.

Also, it is more interesting to study this dependence in weak inversion where possible screening of the Coulomb scattering can be explored by analyzing the dependence of the mobility on both trap density and inversion charge density. However, high channel resistance in weak inversion sets a lower limit on the measurement time. The shortest devices in this work have a gate length of 5 μm. Reducing the gate length to 1 μm lowers the time constant of the transistor by a factor of 25 and easily extends the analysis performed in this work to a time scale of 100 ns and less. Nonetheless, the data presented in this section clearly identify the time scale of the two charge trapping regimes and indicate the fact that mobility only depends on the density of fast traps.

4.7 Prospects of Germanium MOSFETs

The correlation between charge trapping and mobility degradation explored in the previous section can be employed to estimate the expected electron mobility in the absence of charge trapping for this gate stack. Extrapolation of the data in Figure 4-13 suggests that electron mobility as high as 200 $cm^2/V.s$ at lower gate voltages and about 100 $cm^2/V.s$ at the highest gate voltage (4 V) is expected if density of fast traps is reduced to about 1×10^{11} cm^{-2}, which roughly correspond to the state-of-the-art for high-κ dielectrics on silicon. To better estimate the upper limits of electron mobility in Ge MOSFETs, the dependence of mobility on the density of trapped charges in phosphorus-passivated devices is analyzed.

Figure 4-14 (a) shows the pulse-width dependence of the density of trapped charges for the Ge NMOSFETs with different doses of phosphorus passivation implants. Similar to Figure 4-12 (a) two distinct regimes of fast and slow trapping are visible, one at time scales less than about 20 μs and one at longer times. Phosphorus passivation reduces the density of fast traps significantly, while the density of slow traps (obtained by extrapolating the temporal variation of the fast traps to larger pulse widths and subtracting from the total trap density) is almost unchanged. This is consistent with the assumption that fast traps are located at the Ge-dielectric interface and hence



(a) (b)

Figure 4-14: Pulse-width dependence of the density of trapped charges (a) and drain current (b) for Ge NMOSFETs with different doses of phosphorus passivation implant and at a constant inversion charge density of about 1.7×10^{13} cm^{-2} at the beginning of the pulse (corresponding to a pulse amplitude of about 3 V).

phosphorus atoms at the interface are able to passivate them to some extent. Slow traps, located inside the high-κ dielectric are not affected by phosphorus atoms.

Figure 4-14 (b) shows the pulse-width dependence of the drain current for the same transistors. The temporal degradation of the current has been reduced significantly for the devices that received the phosphorus implant.

Finally, Figure 4-15 compares the dependence of electron mobility on the density of trapped charges for transistors that were implanted with different doses of phosphorus. A common dependence of the mobility on the density of fast traps is observed. Extrapolating this dependence to smaller trap densities, and assuming that the mobility would still be limited by Coulomb scattering, suggests that electron mobility as high as 100 cm^2/V.s is expected if the density of fast traps is lowered to about 10^{11} cm^{-2}. While this shows about 75% improvement over the highest mobility achieved in this work, the numbers are still far below what is obtained in silicon MOSFETs and at the same effective electric fields.

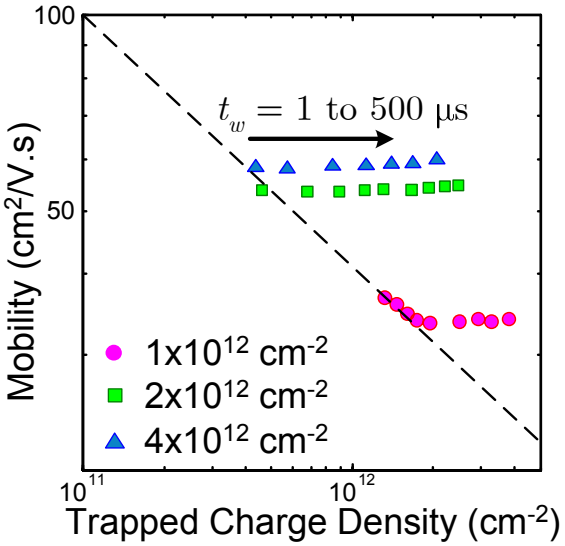


Figure 4-15: The dependence of the electron mobility on trapped charge density for transistors that received different dose of P implant and at an inversion charge density of about 1.7×10^{13} cm^{-2} at the beginning of the pulse. Only fast traps, with a time constant of less than about 10 μs affect the mobility. Extrapolation of the data suggests that if the density of fast traps is reduced to about 1×10^{11} cm^{-2}, mobility will increase by about 40% compared to the devices with the highest dose of P implant explored in this work. At an effective electric field of about 0.64 MV/cm, this is still 2× lower than universal electron mobility in silicon.

4.8 Conclusions

Pulsed $I-V$ and $Q-V$ measurement were performed to characterize near intrinsic transport properties in Ge-channel NMOSFETs. Pulsed measurements showed that the actual carrier mobility is at least twice what is inferred from DC measurements for Ge NFETs. With phosphorus passivation the difference between DC and pulsed measurements was reduced to about 20%, despite the fact that effects of charge trapping are still visible in these devices.

To better understand the dependence of carrier transport on charge trapping a method to directly measure the inversion charge density by integrating the S/D current was proposed. The density of trapped charges was measured as the difference between the inversion charge density at the beginning and end of the pulses applied to the gate. Analysis of temporal variation of trapped charge density revealed that two distinct regimes of fast and slow charge trapping are present. Both mechanisms showed a logarithmic dependence on the pulse width, as observed in earlier literature of charge-pumping studies of Si MOSFETs with high-κ dielectrics. The correlation between mobility and density of trapped charges was studied and it was shown that the mobility depends only on the density of fast traps. To our knowledge this is the first investigation in which the impact of the fast and slow traps on electron mobility in general and in Ge in particular has been separated.

It was demonstrated that phosphorus doping near the Ge-dielectric interface [171] has a passivating effect in that the density of fast traps is reduced, leading to increased electron mobility, while the density of slow traps is essentially unchanged. Extrapolation of the mobility-defect relationship to lower densities of trapped charges gives an upper limit of the available mobility with the present gate stack if the density of the fast traps can be further reduced. However, this analysis demonstrates that the expected mobility is still far below what is obtained in Si MOSFETs. Further investigations are needed to analyze other mechanisms that might be responsible for poor electron mobility in Ge MOSFETs and thereby optimize the gate stack by suppressing these mechanisms. Appendix D presents an interesting approach that relies on

galvanomagnetic measurements to measure the mobility in the inversion layer without the need to determine the inversion charge to open up new possibilities to explore the carrier transport in MOSFETs with new channel materials.

Chapter 5

Summary and Future Work

This chapter gives a summary of the thesis and the main conclusions. Technological implications of the work are also discussed. Major contributions are then listed, followed by suggestions for future work.

5.1 Thesis Summary

This thesis explored some of the options to enhance carrier transport in the MOSFET channel and thereby improve transistor performance. Chapter 2 provided an in depth analysis of transistor performance, defined as the intrinsic MOSFET switching speed, and explored its dependence on different device parameters. A roadmapping exercise was presented and it was shown that new channel materials are needed to lever carrier velocity beyond what is achieved with uniaxially strained silicon, along with dramatic reduction in the device parasitics in order to extend the historical performance scaling trend in the future technology nodes. Such innovations are needed as early as the 32-nm node, to avoid the otherwise counter-scaling of the MOSFET performance.

Chapter 3 analyzed possibilities to improve carrier transport by introducing new channel materials and/or using different strain configurations. It was demonstrated that in uniaxially strained silicon, the virtual source velocity depends more strongly on mobility than previously believed. The modulation of the effective mass under uniaxial strain was shown to be responsible for this strong dependence. It was shown

that the present band structure models are unable to quantitatively reproduce the modulation of the electron effective mass deduced from experimental data, and hence are inadequate to predict the limits of NMOS performance enhancement offered by uniaxial tensile strain. Analysis of valence band structure in uniaxially strained silicon revealed that hole ballistic velocity enhancement is limited to about 100% despite the fact that mobility enhancement as high as 4× has been demonstrated experimentally and predicted theoretically. Supported by analysis of published data, this observation sets a limit on the level of performance enhancement that is possible by uniaxial compressive strain.

Other material systems were explored to see whether they offer performance improvement over silicon. III-V semiconductors were shown to be seriously limited by their small quantization effective mass, which degrades the available inversion charge at a given voltage overdrive. Germanium is in-principle especially attractive as it has enhanced transport properties for both electrons and holes. To avoid mobility degradation due to carrier confinement as well as $L - \Delta$ interband scattering, and to achieve higher ballistic velocity, it was shown that (111) wafer orientation should be used for NFETs. Uniaxial compressive strain was shown to be essential for Ge PFETs in order to beat the velocity achieved in today's uniaxially strained Si transistors.

Despite the attractiveness of Ge as a new channel material, electron mobility measured in Ge MOSFETs is still far below what is expected theoretically. To investigate causes of poor electron transport, pulsed $I - V$ measurements were performed. The results reported in Chapter 4 showed that in the absence of charge trapping, the "intrinsic" electron mobility could be twice what is extracted from DC measurements. A new method was proposed to directly measure the inversion charge density and used to investigate the charge trapping in Ge NFETs. Two regimes of fast and slow charge trapping were observed and it was demonstrated that electron mobility strongly depends on the density of fast traps. The mechanism behind mobility improvement in phosphorus-passivated Ge NFETs was studied by analyzing the mobility-trap density relationship and it was shown that P-passivation reduces the density of fast traps located at the Ge-dielectric interface. Extrapolation of the mobility-trap relationship to

lower densities of trapped charges gives an upper limit on the available mobility with the present gate stack if the density of the fast traps is reduced further. However, the analysis presented in Chapter 4 demonstrates that the expected mobility is still far below what is obtained in Si MOSFETs. Further investigations are needed to analyze other mechanisms that might be responsible for poor electron mobility in Ge MOSFETs and thereby optimize the gate stack by suppressing these mechanisms.

5.2 Technological Implications

The analytical expression given in Chapter 2 for the intrinsic MOSFET delay provides an easy way to explore the impact of different device parameters on the performance. The methodology outlined to extract the virtual source velocity is a handy tool to estimate the effectiveness of different channel engineering methods. Unlike the current practice of normalizing the drive current or the saturation transconductance with the inversion capacitance, it provides an intrinsic measure of the carrier transport inside the channel.

From the discussion in Chapter 2 it is clear that uniaxially strained silicon is unable to provide the velocity needed to meet the target delay at the 32-nm technology node and beyond. Reducing the parasitic capacitance and the subthreshold swing are required to continue the historical trend of MOSFET performance scaling in the future nodes. In fact, these are more feasible than increasing the carrier velocity by either increasing the strain level or by employing new channel materials.

There is no foreseeable limit for increased electron velocity in unixially strained Si. Another 40-50% increase in NFET drive current should be possible if the strain level is doubled. Although this might seem challenging, it could be quite possible by proper substrate engineering. Given that higher dielectric constant of the stressor liners leads to higher parasitic capacitance, methods such as strain memorization techniques or preferential relaxation of biaxial strain should be investigated more seriously.

Further increase of the strain level beyond about 1% is not necessary for improved PFET performance as it does not translate to higher velocity. Optimization of em-

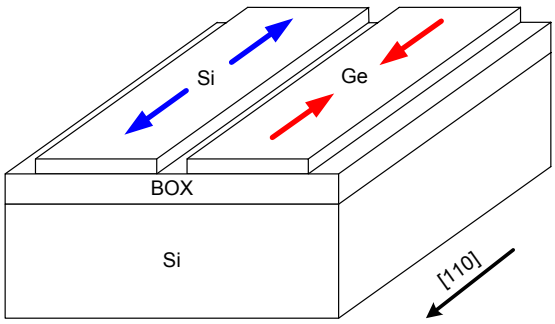


Figure 5-1: The ultimate substrate for high-performance CMOS with ultrathin semiconductor-on-insulator structure that consists of uniaxially strained Si and Ge (or SiGe) stripes for NMOS and PMOS transistors, respectively. To avoid increased parasitic capacitance due to stressor liners, one possible realization of this structure is to start from a biaxially strained heterostructure-on-insulator wafer and preferentially relax the strain. Preferential etching of the semiconductor layers can then be used to obtain the desired channel material.

bedded SiGe process to assert the required strain without the need of stress liners is advantegeous as it lowers the series resistance with minimal penalty in terms of parasitic capacitance. However, for SOI structures, where there is not enough depth for SiGe material, alternative methods should be sought. (110) wafer orientation might be able to offer the required hole velocity at lower strain levels.

Exotic channel materials such as Ge and III-V semiconductors are far from being manufacturable solutions. Moreover, theoretical studies show that III-V semiconductors investigated so far do not offer significant advantage over uniaxially strained Si. Electron velocity in Ge with (111) wafer orientation could be significantly higher than today's uniaxially strained Si. However, efforts needed to develop a manufacturable process for Ge MOSFET might be well invested in better substrate engineering to achieve higher strain levels in Si or reducing the parasitic capacitance. Without uniaxial compressive strain, Ge-channel MOSFETs do not offer any advantage over state-of-the-art PFETs. However, with uniaxial strain applied to Ge, astonishingly high hole velocity is expected. One could thus envision an ideal substrate as a semiconductor-on-insulator structure with Si and Ge strips under uniaxial tensile and compressive strain, respectively as shown schematically in Figure 5-1.

5.3 Contributions

The contributions of this work can be divided into three main subjects: (1) Study of MOSFET performance scaling, (2) theoretical exploration of methods to enhance carrier transport in deeply scaled MOSFETs, and (3) characterization of Ge-channel MOSFETs.

5.3.1 Study of MOSFET Performance Scaling

1. An analytical expression for intrinsic MOSFET delay was developed and used to benchmark CMOS technologies.

2. A methodology to estimate virtual source velocity from literature data was presented and the historical trend of velocity scaling was depicted.

3. A case study of the MOSFET performance in 32-nm nodes was presented and the impact of different device parameters was explored.

4. The severity of increased parasitic capacitance in future technology nodes and the associated performance loss were emphasized.

5.3.2 Theoretical Exploration of Methods for Enhanced Carrier Transport

1. The interplay of the ballistic velocity, the carrier mean free path, and the critical length of scattering, and the dependence of these parameters on the carrier effective mass were comprehensively summarized for the first time.

2. It was demonstrated that in unixially strained Si MOSFETs carrier velocity depends on mobility more strongly than what was previously believed.

3. Prospects of Ge NFETs were studied and the (111) wafer orientation was identified as the best option that offers up to 80% higher ballistic current compared to relaxed Si (approximately 40% compared to state-of-the-art uniaxially strained Si).

4. III-V semiconductors were shown to suffer from extremely small quantization effective mass that limits the available inversion charge at a given gate voltage overdrive.

5. Hole ballistic velocity enhancement in uniaxially strained Si was shown to be limited to about a factor of 2, despite the fact that mobility enhancement up to a factor of 4-5 has been theoretically predicted and experimentally observed.

6. It was demonstrated that Ge does not offer any benefit over today's uniaxially strained Si for improved PMOS performance unless uniaxial compressive strain is also applied to Ge-channel MOSFETs.

5.3.3 Characterization of Ge-Channel MOSFETs

1. Pulsed $I-V$ measurements were performed to characterize near intrinsic carrier mobility in Ge-channel MOSFETs for the first time.

2. A method was introduced to directly measure the inversion charge from pulsed measurements.

3. The density of trapped charges in Ge-channel MOSFETs was measured and two regimes of fast and slow charge trapping were identified.

4. Mobility was shown to only depend on the density of fast traps.

5. Phosphorus surface passivation was demonstrated to only reduce the density of fast traps and thereby improve electron mobility in Ge-channel MOSFETs.

5.4 Suggestions for Future Work

The work presented in this thesis can be extended in many directions:

- The performance metric proposed in Chapter 2 can be easily used to analyze power-delay relationship and explore the prospects of performance scaling in power- and power density-limited designs of the future nodes.

- Calibration of the band structure models with experimental data is required, coupled with self-consistent device simulations to quantitatively identify the limits of the performance improvement offered by strained silicon.

- Experimental data are still required to explore the strain-dependence of the carrier transport in long and short channel Ge MOSFETs.

- The methods presented in Chapter 4 can be used to analyze carrier transport in Ge and III-V MOSFET. Devices with shorter gate length are required to investigate charge trapping at smaller time scales.

- Pulsed measurements can be performed at low temperatures to decouple different carrier scattering mechanisms.

Appendix A

Derivation of the MOSFET Performance Metric and Effective Fringing Capacitance

The width-normalized charge transfered to or from the gate during switching is equal to the inversion charge plus the charge due to overlap and fringing capacitances, with the Miller effect taken into account for the drain side [52]:

$$\begin{aligned} \Delta Q/W &= C'_{\text{inv}} L_{eff}(V_{DD} - V_{T0}) \\ &+ 3(C'_{ox} l_{ov} + C_{of})V_{DD} + 2C_{if}V_{T0}, \end{aligned} \tag{A.1}$$

where L_{eff} is the effective channel length, V_{T0} is the linear threshold voltage, equal to $V_T + \delta V_{DD}$, C'_{ox} is the oxide capacitance per unit area in the overlap region, which is assumed to be equal to C'_{inv}, l_{ov} is the overlap length per side, and C_{if} and C_{of} are the inner and outer fringing capacitances per unit width, respectively. The Miller effect is taken into account for the overlap and outer fringing capacitances, whereas the inner fringing capacitance is only considered when the transistor is off [52]. Noting that

$L_G = L_{eff} + 2l_{ov}$, the total charge is equal to:

$$\begin{aligned} \Delta Q/W &= C'_{\text{inv}} L_G (V_{DD} - V_{T0}) + C'_{\text{inv}} l_{ov} (V_{DD} + 2V_{T0}) \\ &+ (3C_{of} + 2C_{if}) V_{T0}. \end{aligned} \tag{A.2}$$

With typical values of 0.06, 0.08, and 0.16 fF/μm for $C'_{\text{inv}} l_{ov}$, C_{of}, and C_{if}, respectively [52], and assuming that $V_{T0} \approx 0.4 V_{DD}$, the last two terms can be lumped into $C^*_f V_{DD}$ with an effective fringing capacitance, C^*_f, of about 0.5 fF/μm. Numerical values obtained from simulation on a typical state-of-the-art MOSFET are slightly different: 0.05, 0.13, and 0.08 F/μm for $C'_{\text{inv}} l_{ov}$, C_{of}, and C_{if}, respectively. This gives a value of 0.56 fF/μm for C^*_f, which corresponds to a total "overlap" capacitance of 0.26 fF/μm which is typical of what is reported in the literature.

The above numbers are obtained in the absence of raised source/drain and when contact plugs are far away from the gate contact. In that case the outer fringing capacitance contains three parts, C_{side}, the fringing capacitance between the gate electrode and planar portion of S/D (equivalent to C_{of} for isoplanar structure), C_{pp}, the parallel plate capacitance between the gate and S/D plugs, and C_{top}, the fringing capacitance from the top of the gate electrode to S/D plugs [179]:

$$C_{of} = C_{side} + C_{pp} + C_{top}, \tag{A.3}$$

where

$$C_{side}/W = \frac{0.8\epsilon_{sp}}{\pi} \ln \left[(\alpha - 1) \left(\frac{\alpha}{\alpha - 1} \right)^{\alpha} \right], \tag{A.4}$$

with $\alpha = (t_{sp}/t_{ox})^2$.

$$C_{pp} = 1.5\epsilon_{sp} n_c W_c \left(\frac{t_{ox} + t_{\text{poly}}}{t_{sp}} - 0.55 \right), \tag{A.5}$$

and

$$C_{top} = \frac{1.6\epsilon_{top} n_c W_c}{\pi} \ln \left(1 + \frac{L_G}{2t_{sp}} \right), \tag{A.6}$$

where ϵ_{sp} is the dielectric permittivity of the spacer, t_{sp} is the spacer thickness, t_{poly}

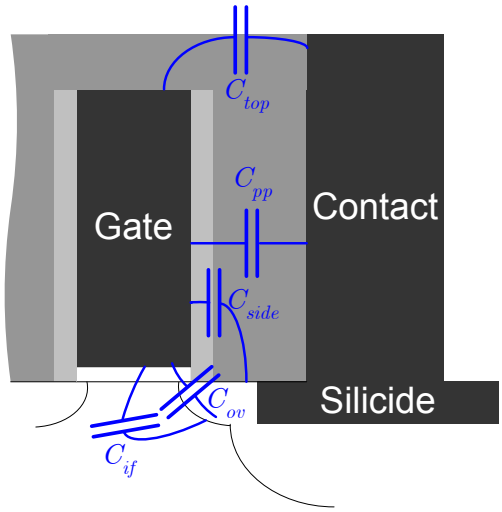


Figure A-1: Illustration of different components of the parasitic capacitance.

is the gate electrode height, t_{ox} is the physical thickness of the gate oxide, and ϵ_{top} is the permittivity of the top layer, a low-temperature oxide for planarization or nitride layers to apply strain to the channel, n_c is the number of contact plugs, and W_c is the contact width and spacing. For a wide transistor with maximum number of contact plugs $n_c W_c = W/2$. These additional capacitance components are schematically shown in Figure A-1. The effective parasitic capacitance increases with technology scaling, as shown in Figure 2-9(a).

Appendix B

Band Structure Calculation in Strained Semiconductors

Applying mechanical strain to the semiconductor distorts its lattice structure. Under uniaxial strain in the [001] direction, the lattice will have an orthorhombic structure, with strain tensor given by:

$$\epsilon_{001} = \begin{bmatrix} \epsilon_{xx} & 0 & 0 \\ 0 & \epsilon_{xx} & 0 \\ 0 & 0 & \epsilon_{zz} \end{bmatrix}, \tag{B.1}$$

where $\epsilon_{xx} = \epsilon$ and $\epsilon_{zz} = -2C_{12}/C_{11}\epsilon$, and C_i's are the elastic stiffness constants. A uniaxial strain in the [110] direction, however, results in a monoclinic crystal. Furthermore, the strain-stress symmetry in the $x - y$ plane breaks up; If mechanical strain is applied in the (110) direction $\epsilon_{xx} = \epsilon_{yy}$, while $\sigma_{xx} \neq \sigma_{yy}$.

$$\epsilon_{110} = \begin{bmatrix} \epsilon_{xx} & \epsilon_{xy} & 0 \\ \epsilon_{xy} & \epsilon_{xx} & 0 \\ 0 & 0 & \epsilon_{zz} \end{bmatrix}, \tag{B.2}$$

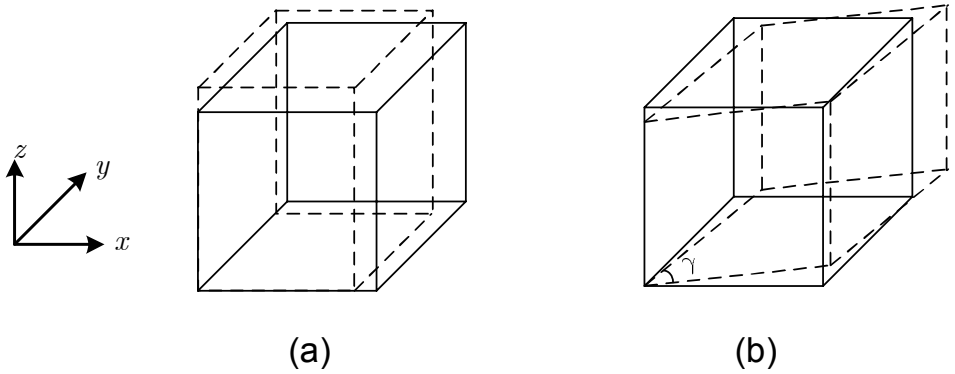


Figure B-1: Distortion of the lattice under (a) uniaxial tensile strain in the (001) direction and (b) uniaxial compressive strain in the (110) direction.

where

$$\begin{aligned} \epsilon_{xx} &= \frac{2C_{44} - C_{12}}{2C_{44} + C_{11} + C_{12}}\epsilon \\ \epsilon_{xy} &= \frac{-(C_{11} + 2C_{12})}{2C_{44} + C_{11} + C_{12}}\epsilon \\ \epsilon_{zz} &= \epsilon \end{aligned}$$

Finally, for a uniaxial strain in the (111) direction, the strain tensor is:

$$\epsilon_{111} = \begin{bmatrix} \epsilon_{xx} & \epsilon_{xy} & \epsilon_{xy} \\ \epsilon_{xy} & \epsilon_{xx} & \epsilon_{xy} \\ \epsilon_{xy} & \epsilon_{xy} & \epsilon_{xx} \end{bmatrix}, \tag{B.3}$$

where

$$\begin{aligned} \epsilon_{xx} &= \frac{4C_{44}}{4C_{44} + C_{11} + C_{12}}\epsilon \\ \epsilon_{xy} &= \frac{-(C_{11} + 2C_{12})}{4C_{44} + C_{11} + C_{12}}\epsilon \end{aligned}$$

The above treatment of the strain only describes the macroscopic picture. In the atomic scale, and when the strain is applied along the (110) or (111) direction, there is an additional (Kleinman's) "internal displacement" between the atoms of the two

Table B.1: Elastic stiffness constants (10^{11} dyn/cm) and internal displacement parameter in Si and Ge.

	C_{11}	C_{12}	C_{44}	ζ
Si	16.577	6.393	7.962	0.53
Ge	12.40	4.13	6.83	0.44

sublattices.

$$\mathbf{R}_i^A = (\mathbf{1} + \epsilon)\mathbf{R}_{i,0}^A \tag{B.4}$$

For strain in the (001) direction:

$$\mathbf{R}_i^B = (\mathbf{1} + \epsilon)\mathbf{R}_{i,0}^B \tag{B.5}$$

(110) direction:

$$\mathbf{R}_i^B = (\mathbf{1} + \epsilon)\mathbf{R}_{i,0}^B - \frac{a_0}{2}\epsilon_{xy}\zeta \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \tag{B.6}$$

and (111) direction:

$$\mathbf{R}_i^B = (\mathbf{1} + \epsilon)\mathbf{R}_{i,0}^B - \frac{a_0}{2}\epsilon_{xy}\zeta \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \tag{B.7}$$

where $\mathbf{R}_i^A$ and $\mathbf{R}_i^B$ are the atom's position in the sublattice A and B, respectively, a_0 is the unstrained lattice constant, and ζ is the internal displacement parameter varying from 0 to 1 to minimize the total energy of the lattice under strain. Unlike elastic stiffness constants, which can be measured experimentally, the internal displacement parameter is often calculated from *ab initio* simulations. Representing values are given in Table B.1.

Appendix C

Hole Effective Mass Anisotropy in Si and Ge

Starting from the 6×6 $k.p$ Hamiltonian, hole effective mass for the three high-symmetry directions can be calculated analytically. Calculations can be done at two limits of low and high energies [180] as shown in Table C.1.

Luttinger parameters in silicon and germanium are given in Table C.2 and used to calculate numerical values of the effective mass given in Table C.3. This table shows some of the important features of the valence band in Si and Ge. At low energies, LH band is almost symmetric with small effective mass, but it approaches the HH band at higher energies. Symmetric and small effective mass of the LH means that under quantization, this band moves to higher energies very quickly. Of course, accurate treatment of the band structure under quantization requires self-

Table C.1: Valence band effective mass along the high-symmetry directions and in terms of the Luttinger parameters [180].

	[100]	[111]	[110]
$E \ll \Delta_{\mathrm{SO}}$	$\frac{1}{\gamma_1 \pm 2\gamma_2}$	$\frac{1}{\gamma_1 \pm 2\gamma_3}$	$\frac{1}{\gamma_1 \pm \sqrt{\gamma_2^2 + 3\gamma_3^2}}$
$E \gg \Delta_{\mathrm{SO}}$	$\frac{1}{\gamma_1 - 2\gamma_2}$	$\frac{1}{\gamma_1 - 2\gamma_3}$	$\frac{2}{2\gamma_1 - (\gamma_2 + 3\gamma_3) \pm 3(\gamma_2 - \gamma_3)}$

Table C.2: Luttinger parameters in Si and Ge.

	γ_1	γ_2	γ_3
Si	4.21	0.427	1.42
Ge	12.60	3.93	5.39

Table C.3: Hole effective mass in Si and Ge.

		[100]	[111]	[110]
$E \ll \Delta_{SO}$				
	Si	0.20	0.14	0.15
		0.30	0.73	0.58
	Ge	0.049	0.043	0.044
		0.21	0.55	0.40
$E \gg \Delta_{SO}$				
	Si	0.30	0.73	0.30
		0.30	0.73	2.65
	Ge	0.21	0.55	0.21
		0.21	0.55	2.77

consistent simulations. Nonetheless, the HH band is the main contributor to the carrier transport in the inversion layer, especially in Ge. From this perspective, the valence band in germanium is quite similar to that in silicon. As discussed in Chapter 3, this is a major limit for the expected hole velocity enhancement in Ge as compared to Si.

Appendix D

Galvanomagnetic Effects for Characterization of Transport in the Inversion Layer

Galvanomagnetic phenomena, such as Hall effect and transverse magnetoresistance, are powerful tools to study the carrier transport in the inversion layer. With an assumption about the Hall scattering factor, both carrier mobility and concentration can be extracted from Hall measurements. Alternatively, by performing the measurements at the limits of low and high magnetic field, the scattering factor can be extracted as well [181]. Hence, this method is especially attractive in characterizing MOSFETs with high-κ dielectrics, where charge trapping makes the determination of the inversion charge from split-CV measurements ambiguous. The low-field Hall coefficient is given by:

$$R_H = \frac{r_H}{qn}, \tag{D.1}$$

where r_H is the Hall scattering factor, n is the carrier concentration, and q is the electronic charge. Theory predicts that r_H approaches unity in the limit of high magnetic field, so $r_H(B) = R_H(B)/R_H(\infty)$, where B is the magnetic field. The high magnetic field limit is given by $\mu_D B \gg 1$, where μ_D is the drift mobility. In practice, a plot of the $R_H(B)$ can be made to see whether it saturates to its minimum at high

magnetic fields [181]. Once the scattering factor is determined, the carrier density and mobility are easily calculated: $\mu_D = \mu_H / r_H$, where μ_H is the Hall mobility. The procedure to perform Hall measurements is very straightforward. However, it requires special structures, such as van der Pauw structure or Hall bars, which are not standard on a MOSFET layout.

Alternatively, transverse magnetoresistance can be used to measure the mobility (magnetoresistance mobility, μ_M) without the need to any special structure. In the ideal case, where the width of the transistor is much larger than its length, the normalized channel resistance is given by:

$$\frac{R(B)}{R(0)} = 1 + \mu_M^2 B^2. \tag{D.2}$$

For a finite width to length ratio:

$$\frac{R(B)}{R(0)} = \begin{cases} 1 + \mu_M^2 B^2 (1 - 0.543 \dfrac{L}{W}) & \text{for } L/W < 0.3 \\ 1 + \mu_M^2 B^2 (0.543 \dfrac{L}{W}) & \text{for } L/W > 1 \end{cases}$$

where W and L are the channel width and length, respectively [182]. The former is usually the case for transistor structures, while the latter is the case for Hall bars. Ref. [182] tabulates the geometrical correction factor the L/W values that lie between the above two limits.

In general the effective mobility, Hall mobility, and magnetoresistance mobility

are given by:

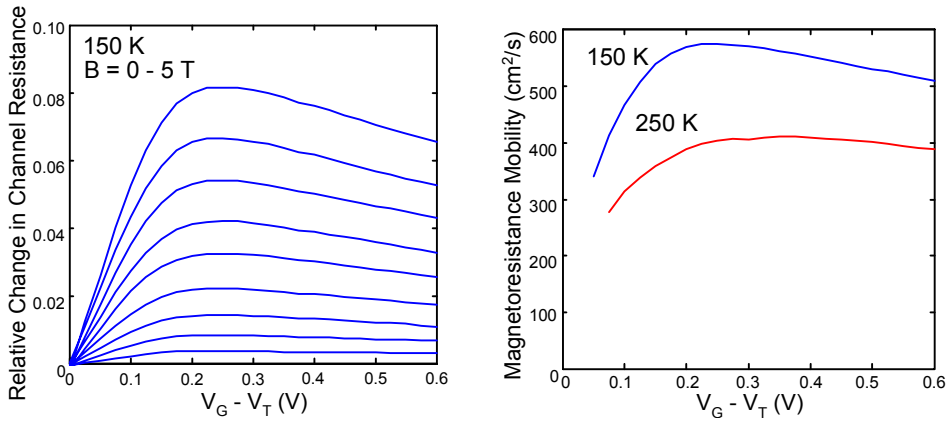
$$\begin{aligned}
\mu_{\text{eff}} &= \frac{q\langle\tau_m\rangle}{m^*} = \frac{q}{m^*}\frac{\displaystyle\int_0^\infty -\frac{\partial f}{\partial E}ED(E)\tau_m(E)dE}{\displaystyle\int_0^\infty f(E)D(E)dE} \\
\mu_H &= \frac{q\langle\tau_m^2\rangle}{m^*\langle\tau_m\rangle} = \frac{q}{m^*}\frac{\displaystyle\int_0^\infty -\frac{\partial f}{\partial E}ED(E)\tau_m(E)^2dE}{\displaystyle\int_0^\infty -\frac{\partial f}{\partial E}ED(E)\tau_m(E)dE} \\
\mu_{MR} &= \frac{q}{m^*}\sqrt{\frac{\langle\tau_m^3\rangle}{\langle\tau_m\rangle}} = \frac{q}{m^*}\sqrt{\frac{\displaystyle\int_0^\infty -\frac{\partial f}{\partial E}ED(E)\tau_m(E)^3dE}{\displaystyle\int_0^\infty -\frac{\partial f}{\partial E}ED(E)\tau_m(E)dE}}
\end{aligned} \tag{D.3}$$

If we assume that the momentum relaxation time has a power-law dependence on the carrier energy, $\tau_m \sim E^s$, with non-degenerate carrier statistics, the Hall scattering factor and magnetoresistance ratio are easily obtained:

$$\begin{aligned}
r_H &= \frac{\Gamma(2s+2)\Gamma(2)}{\Gamma(s+2)^2} \\
r_M &= \sqrt{\frac{\Gamma(3s+2)}{\Gamma(s+2)^3}},
\end{aligned} \tag{D.4}$$

where $\Gamma(.)$ is the gamma function. In reality, many subbands contribute to carrier transport and degenerate statistics should be used at high inversion charge density. Monte Carlo simulations are then necessary to determine r_H and r_M [183].

Figure D-1 shows the results of magnetoresistance measurements performed on Ge PFETs. The magnetic field was applied normal to the transistor plane and was swept up to 5 T. Electrical measurements were performed on circular transistors that satisfy $W \gg L$. Figure D-1 shows the relative change in the channel resistance as a function of the gate voltage overdrive. Magnetoresistance mobility is extracted by fitting the quadratic expression $R(B)/R(0) = 1+\mu_M^2 B^2$ and the results are shown in Figure D-1. The magnetoresistance mobility reported in Figure D-1 is approximately three times the effective mobility measured on the same samples [138]. This difference cannot be



(a) (b)

Figure D-1: Sample magnetoresistance measurements performed on Ge PMOS transistors. (a) Relative change in the channel resistance as a function of the gate voltage for different magnetic fields applied normal to the transistor plane and at 150 K. (b) Extracted magnetoresistance mobility as a function of the gate voltage and at two different temperatures.

explained by the theory of magnetoreistance, which predicts that r_M is roughly 1.8 for mobility limited by Coulomb scattering. More investigations are thus needed to understand carrier transport in Ge inversion layers.

Bibliography

[1] G. Moore. Cramming more components onto integrated circuits. *Electronics Magazine*, 38:114–117, 1965.

[2] 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*, 9:256–268, 1974.

[3] G. Van den Bosch, S. Demuynck, Zs. Tökei, G. Beyer, M. Van Hove, and G. Groesenken. Impact of copper contacts on CMOS front-end yield and reliability. In *IEDM Tech. Dig.*, pages 93–96, 2006.

[4] M. Bohr, S.S. Ahmed, S.U. Ahmed, M. Bost, T. Ghani, J. Greason, R. Hainsey, C. Jan, P. Packan, S. Sivakumar, S. Thompson, J. Tsai, and S. Yang. A high performance 0.25 μm logic technology optimized for 1.8V operation. In *IEDM Tech. Dig.*, pages 847–850, 1996.

[5] S. Yang, S. Ahmed, B. Arcot, R. Arghavani, P. Bai, S. Chambers, P. Charvat, R. Cotner, R. Gasser, T. Ghani, M. Hussein, C. Jan, C. Kardas, J. Maiz, P. McGregor, B. McIntyre, P. Nguyen, P. Packan, I. Post, S. Sivakumar, J. Steigerwald, M. Taylor, B. Tufts, S. Tyagi, and M. Bohr. A high performance 180 nm generation logic technology. In *IEDM Tech. Dig.*, pages 197–200, 1998.

[6] T. Ghani, S. Ahmed, P. Aminzadeh, J. Bielefeld, P. Charvat, C. Chu, M. Harper, P. Jacob, C. Jan, J. Kavalieros, C. Kenyon, R. Nagisetty, P. Packan, J. Sebastian, M. Taylor, J. Tsai, S. Tyagi, S. Yang, and M Bohr. 100 nm gate

length high performance/low power CMOS transistor structure. In *IEDM Tech. Dig.*, pages 415–418, 1999.

[7] S. Tyagi, M. Alavi, R. Bigwood, T. Bramblett, J. Brandenburg, W. Chen, B. Crew, M. Hussein, P. Jacob, C. Kenyon, C. Lo, B. McIntyre, Z. Ma, P. Moon, P. Nguyen, L. Rumaner, R. Schweinfurth, S. Sivakumar, M. Stettler, S. Thompson, B. Tufts, J. Xu, S. Yang, and M. Bohr. A 130 nm generation logic technology featuring 70 nm transistors, dual vt transistors and 6 layers of Cu interconnects. In *IEDM Tech. Dig.*, pages 567–570, 2000.

[8] S. Thompson, M. Alavi, R. Arghavani, A. Brand, R. Bigwood, J. Brandenburg, B. Crew, V. Dubin, M. Hussein, P. Jacob, C. Kenyon, E. Lee, B. Mcintyre, Z. Ma, P. Moon, P. Nguyen, M. Prince, R. Schweinfurth, S. Sivakumar, P. Smith, M. Stettler, S. Tyagi, M. Wei, J. Xu, S. Yang, and M Bohr. An enhanced 130 nm generation logic technology featuring 60 nm transistors optimized for high performance and low power at 0.7 - 1.4 V. In *IEDM Tech. Dig.*, pages 257–260, 2001.

[9] S. Thompson, N. Anand, M. Armstrong, C. Auth, B. Arcot, M. Alavi, P. Bai, J. Bielefeld, R. Bigwood, J. Brandenburg, M. Buehler, S. Cea, V. Chikarmane, C. Choi, R. Frankovic, T. Ghani, G. Glass, W. Han, T. Hoffmann, M. Hussein, P. Jacob, A. Jain, C. Jan, S. Joshi, C. Kenyon, J. Klaus, S. Klopcic, J. Luce, Z. Ma, B. Mcintyre, K. Mistry, A. Murthy, P. Nguyen, H. Pearson, T. Sandford, R. Schweinfurth, R. Shaheed, S. Sivakumar, M. Taylor, B. Tufts, C. Wallace, P. Wang, C. Weber, and M. Bohr. A 90 nm logic technology featuring 50 nm strained silicon channel transistors, 7 layers of Cu interconnects, low k ILD, and 1 μm^2 SRAM cell. In *IEDM Tech. Dig.*, pages 61–64, 2002.

[10] T. Ghani, M. Armstrong, C. Auth, M. Bost, P. Charvat, G. Glass, T. Hoffmann, K. Johnson, C. Kenyon, J. Klaus, B. McIntyre, K. Mistry, A. Murthy, J. Sandford, M. Silberstein, S. Sivakumar, P. Smith, K. Zawadzki, S. Thompson, and M. Bohr. A 90nm high volume manufacturing logic technology featuring novel

45nm gate length strained silicon CMOS transistors. In *IEDM Tech. Dig.*, pages 978–981, 2003.

[11] P. Bai, C. Auth, S. Balakrishnan, M. Bost, R. Brain, V. Chikarmane, R. Heussner, M. Hussein, J. Hwang, D. Ingerly, R. James, J. Jeong, C. Kenyon, E. Lee, S.-H. Lee, N. Lindert, M. Liu, Z. Ma, T. Marieb, A. Murthy, R. Nagisetty, S. Natarajan, J. Neirynck, A. Ott, C. Parker, J. Sebastian, R. Shaheed, S. Sivakumar, J. Steigerwald, S. Tyagi, C. Weber, B. Woolery, A. Yeoh, K. Zhang, and M. Bohr. A 65nm logic technology featuring 35nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, low-k ILD and 0.57 μm^2 SRAM cell. In *IEDM Tech. Dig.*, pages 657–660, 2004.

[12] S. Tyagi, C. Auth, P. Bai, G. Curello, H. Deshpande, S. Gannavaram, O. Golonzka, R. Heussner, R. James, C. Kenyon, S.-H. Lee, N. Lindert, M. Liu, R. Nagisetty, S. Natarajan, C. Parker, J. Sebastian, B. Sell, S. Sivakumar, A.S. Amour, and K Tone. An advanced low power, high performance, strained channel 65nm technology. In *IEDM Tech. Dig.*, pages 1070–1072, 2005.

[13] K. Mistry, C. Allen, C. Auth, B. Beattie, D. Bergstrom, M. Bost, M. Brazier, M. Buehler, A. Cappellani, R. Chau, C.-H. Choi, G. Ding, K. Fischer, T. Ghani, R. Grover, W. Han, D. Hanken, M. Hattendorf, J. He, J. Hicks, R. Huessner, D. Ingerly, P. Jain, R. James, L. Jong, S. Joshi, C. Kenyon, K. Kuhn, K. Lee, H. Liu, J. Maiz, B. McIntyre, P. Moon, J. Neirynck, S. Pae, C. Parker, D. Parson, C. Prasad, L. Pipes, M. Prince, P. Ranade, T. Reynolds, J. Sanford, L. Shifren, J. Sebastian, J. Seiple, D. Simon, S. Sivakumar, P. Smith, c. Thomas, T. Troeger, P. Vandervoon, S. Williams, and K. Zawadzki. A 45nm logic technology with high-k+metal gate transistors, strained silicon, 9 Cu interconnect layers, 193nm dry patterning, and 100% Pb-free packaging. In *IEDM Tech. Dig.*, pages 247–250, 2007.

[14] F. Boeuf, F. Arnaud, C. Boccaccio, F. Salvetti, J. Todeschini, L. Pain, M. Jurdit, S. Manakli, B. Icard, N. Planes, N. Gierczynski, S. Denorme, B. Borot,

C. Ortolland, B. Duriez, B. Tavel, P. Gouraud, M. Broekaart, V. Dejonghe, P. Brun, F. Guyader, P. Morini, C. Reddy, M. Aminpur, C. Laviron, S. Smith, J.P. Jacquemin, M. Mellier, F. Andre, N. Bicais-Lepinay, S. Jullian, J. Bustos, and T. Skotnicki. $0.248\mu m^2$ and $0.334\mu m^2$ conventional bulk 6T-SRAM bit-cells for 45nm node low cost - general purpose applications. In *Symp. VLSI Tech.*, pages 130–131, 2005.

[15] B. Davari, R. H. Dennard, and G. G. Shahidi. CMOS scaling for high performance and low power- the next ten years. *Proc. IEEE*, 83(4):594–606, 1995.

[16] S. Borkar. Circuit techniques for subthreshold leakage avoidance, control and tolerance. In *IEDM Tech. Dig.*, pages 421–424, 2004.

[17] J.L. Hoyt, H.M. Nayfeh, S. Eguchi, I. Åberg, G. Xia, T. Drake, E.A. Fitzgerald, and D.A. Antoniadis. Strained silicon MOSFET technology. In *IEDM Tech. Dig.*, pages 23–26, 2002.

[18] I. Lauer, T.A. Langdo, Z.-Y. Cheng, J.G. Fiorenza, G. Braithwaite, M.T. Currie, C.W. Leitz, A. Lochtefeld, H. Badawi, M.T. Bulsara, M. Somerville, and D.A. Antoniadis. Fully depleted n-MOSFETs on supercritical thickness strained SOI. *IEEE Electron Device Lett.*, 25(2):83–85, 2004.

[19] K. Rim, K. Chan, L. Shi, D. Boyd, J. Ott, N. Klymko, F. Cardone, L. Tai, S. Koester, M. Cobb, D. Canaperi, B. To, E. Duch, I. Babich, R. Carruthers, P. Saunders, G. Walker, Y. Zhang, M. Steen, and M. Ieong. Fabrication and mobility characteristics of ultra-thin strained Si directly on insulator (SSDOI) MOSFETs. In *IEDM Tech. Dig.*, pages 49–52, 2003.

[20] L. Gomez, I. Åberg, and J. L. Hoyt. Electron transport in strained-silicon directly on insulator ultrathin-body n-MOSFETs with body thickness ranging from 2 to 25 nm. *IEEE Electron Device Lett.*, 28:285–287, 2007.

[21] K. Rim, S. Koester, M. Hargrove, J. Chu, P.M. Mooney, J. Ott, T. Kanarsky, P. Ronsheim, M. Ieong, A. Grill, and H.-S.P. Wong. Strained Si NMOSFETs

for high performance CMOS technology. In *Symp. VLSI Tech.*, pages 59–60, 2001.

[22] I. Åberg, C. Ní Chléirigh, O. O. Olubuyide, X. Duan, and J. L. Hoyt. High electron and hole mobility enhancements in thin-body strained Si/strained SiGe/strained Si heterostructures on insulator. In *IEDM Tech. Dig.*, pages 173–176, 2004.

[23] K. Goto, S. Satoh, H. Ohta, S. Fukuta, T. Yamamoto, T. Mori, Y. Tagawa, T. Sakuma, T. Saiki, Y. Shimamune, A. Katakami, A. Hatada, H. Morioka, Y. Hayami, S. Inagaki, K. Kawamura, Y. Kim, H. Kokura, N. Tamura, N. Horiguchi, M. Kojima, T. Sugii, and K. Hashimoto. Technology booster using strain-enhancing laminated SiN (SELS) for 65nm node HP MPUs. In *IEDM Tech. Dig.*, pages 209–212, 2004.

[24] Z. Luo, A. Steegen, M. Eller, R. Mann, C. Baiocco, P. Nguyen, L. Kim, M. Hoinkis, V. Ku, V. Klee, F. Jamin, P. Wrschka, P. Shafer, W. Lin, S. Fang, A. Ajmera, W. Tan, D. Park, R. Mo, J. Lian, D. Vietzke, C. Coppock, A. Vayshenker, T. Hook, V. Chan, K. Kim, A. Cowley, S. Kim, E. Kaltalioglu, B. Zhang, S. Marokkey, Y. Lin, K. Lee, H. Zhu, M. Weybright, R. Rengarajan, J. Ku, T. Schiml, J. Sudijono, I. Yang, and C. Wann. High performance and low power transistors integrated in 65nm bulk cmos technology. In *IEDM Tech. Dig.*, pages 661–664, 2004.

[25] T. Komoda, A. Oishi, T. Sanuki, K. Kasai, H. Yoshimura, K. Ohno, A. Iwai, M. Saito, F. Matsuoka, N. Nagashima, and T. Noguchi. Mobility improvement for 45nm node by combination of optimized stress and channel orientation design. In *IEDM Tech. Dig.*, pages 217–220, 2004.

[26] M. Shima, T. Ueno, T. Kumise, H. Shido, Y. Sakuma, and S. Nakamura. ⟨100⟩ channel strained-SiGe p-MOSFET with enhanced hole mobility and lower parasitic resistance. In *Symp. VLSI Tech.*, pages 94–95, 2002.

[27] T. Mizuno, N. Sugiyama, T. Tezuka, Y. Moriyama, S. Nakaharai, and S. Takagi. (110)-surface strained-SOI CMOS devices with higher carrier mobility. In *Symp. VLSI Tech.*, pages 97–98, 2003.

[28] M. Yang, M. Ieong, L. Shi, K. Chan, V. Chan, A. Chou, E. Gusev, K. Jenkins, D. Boyd, Y. Ninomiya, D. Pendleton, Y. Surpris, D. Heenan, J. Ott, K. Guarini, C. D'Emic, M. Cobb, P. Mooney, B. To, N. Rovedo, J. Benedict, R. Mo, and H Ng. High performance CMOS fabricated on hybrid substrate with different crystal orientations. In *IEDM Tech. Dig.*, pages 453–456, 2003.

[29] C.D. Sheraw, M. Yang, D.M. Fried, G. Costrini, T. Kanarsky, W.-H. Lee, V. Chan, M.V. Fischetti, J. Holt, L. Black, M. Naeem, S. Panda, L. Economikos, J. Groschopf, A. Kapur, Y. Li, R.T. Mo, A. Bonnoit, D. Degraw, S. Luning, D. Chidambarrao, X. Wang, A. Bryant, D. Brown, C.-Y. Sung, P. Agnello, M. Ieong, S.-F. Huang, X. Chen, and M. Khare. Dual stress liner enhancement in hybrid orientation technology. In *Symp. VLSI Tech.*, pages 12–13, 2005.

[30] B. Yang, K. Nummy, A. Waite, L. Black, H. Gossmann, H. Yin, Y. Liu, B. Kim, S. Narasimha, P. Fisher, M. V. Meer, J. Johnson, D. Chidambarrao, S. D. Kim, C. Sheraw, D. Wehella-gamage, J. Holt, X. Chen, D. Park, C.Y. Sung, D. Schepis, M. Khare, S. Luning, and P. Agnello. Stress dependence and poly-pitch scaling characteristics of (110) PMOS drive current. In *Symp. VLSI Tech.*, pages 126–127, 2007.

[31] P.D. Ye, G.D. Wilk, J. Kwo, B. Yang, H.-J.L. Gossmann, M. Frei, S.N.G. Chu, J.P. Mannaerts, M. Sergent, M. Hong, K.K. Ng, and J. Bude. GaAs MOSFET with oxide gate dielectric grown by atomic layer deposition. *IEEE Electron device Lett.*, 24:209 – 211, 2003.

[32] K. Rajagopalan, R. Droopad, J. Abrokwah, P. Zurcher, P. Fejes, and M. Passlack. 1-μm enhancement mode GaAs n-channel MOSFETs with transconductance exceeding 250 mS/mm. *IEEE Electron Device Lett.*, 28:100–102, 2007.

[33] H. Shang, H. Okorn-Schmidt, K. K. Chan, M. Copel, J. A. Ott, P. M. Kozlowski, Steen, S.A. Cordes, H.-S.P. Wong, E.C. Jones, and W.E. Haensch. High mobility p-channel germanium MOSFETs with a thin Ge oxynitride gate dielectric. In *IEDM Tech. Dig.*, pages 441–444, 2002.

[34] H. Shang, J.O. Chu, X. Wang, P.M. Mooney, K. Lee, J. Ott, K. Rim, K. Chan, K. Guarini, and M. Ieong. Channel design and mobility enhancement in strained germanium buried channel MOSFETs. In *Symp. VLSI Tech.*, pages 204–205, 2004.

[35] H. Shang, K.-L. Lee, P. Kozlowski, C. D'Emic, I. Babich, E. Sikorski, M. Ieong, H.-S.P. Wong, K. Guarini, and W. Haensch. Self-aligned n-channel germanium MOSFETs with a thin Ge oxynitride gate dielectric and tungsten gate. *IEEE Electron Device Lett.*, 25(3):135–137, 2004.

[36] H. Shang, H. Okorn-Schimdt, J. Ott, P. Kozlowski, S. Steen, E.C. Jones, H.-S.P. Wong, and W. Hanesch. Electrical characterization of germanium p-channel MOSFETs. *IEEE Electron Device Lett.*, 24(4):242–244, 2003.

[37] C. O. Chui, H. Kim, D. Chi, B. B. Triplett, P. C. McIntyre, and K. C. Saraswat. A sub-400° C germanium MOSFET technology with high-κ dielectric and metal gate. In *IEDM Tech. Dig.*, pages 437–440, 2002.

[38] A. Ritenour, S. Yu, M.L. Lee, Z. Lu, W. Bai, A. Pitera, E.A. Fitzgerald, D.L. Kwong, and D.A. Antoniadis. Epitaxial strained germanium p-MOSFETs with HfO_2 gate dielectric and TaN gate electrode. In *IEDM Tech. Dig.*, 2003.

[39] M. L. Lee, C. W. Leitz, Z. Cheng, A. J. Pitera, T. Langdo, M. T. Currie, G. Taraschi, E. A. Fitzgerald, and D. A. Antoniadis. Strained Ge channel p-type metaloxidesemiconductor field-effect transistors grown on $Si_{1-x}Ge_x$/Si virtual substrates. *Appl. Phys. Lett.*, 79(20):3344–3346, 2001.

[40] Y. Cao, T. Sato, D. Sylvester, M. Orshansky, and C. Hu. New paradigm of predictive MOSFET and interconnect modeling for early circuit design. In *Proc. Custom Integrated Circuits Conf.*, pages 201–204, 2000.

[41] W. Zhao and Y. Cao. New generation of predictive technology model for sub-45nm early design exploration. *IEEE Trans. Electron Devices*, 53:2816–2823, 2006.

[42] M.H. Na, E.J. Nowak, W. Haensch, and J. Cai. The effective drive current in CMOS inverters. In *IEDM Tech. Dig.*, pages 121–124, 2002.

[43] J. Ding and H.-S. P. Wong. Metrics for performance benchmarking of nanoscale Si and carbon nanotube FETs including device nonidealities. *IEEE Trans. Electron Devices*, 53:1317–1322, 2006.

[44] E. Yoshida, Y. Momiyama, M. Miyamoto, T. Saiki, M. Kojima, S. Satoh, and T. Sugii. Performance boost using a new device design methodology based on characteristic current for low-power CMOS. In *IEDM Tech. Dig.*, pages 195–198, 2006.

[45] D. A. Antoniadis, I. Åberg, C. Ní Chléirigh, O. M. Nayfeh, A. Khakifirooz, and J. L. Hoyt. Continuous MOSFET performance increase with device scaling: the role of strain and channel material innovations. *IBM J. Res. Dev.*, 50(4/5):363–376, 2006.

[46] A. Khakifirooz and D. A. Antoniadis. Transistor performance scaling: The role of virtual source velocity and its mobility dependence. In *IEDM Tech. Dig.*, pages 667–670, 2006.

[47] A. Lochtefeld and D. A. Antoniadis. On experimental determination of carrier velocity in deeply scaled NMOS: how close to the thermal limit? *IEEE Electron Device Lett.*, 22(2):95–97, 2001.

[48] C.-H. Choi, P. R. Chidambaram, R. Khamankar, C. F. Machala, Z. Yu, and R. W. Dutton. Dopant profile and gate geometric effects on polysilicon gate depletion in scaled MOS. *IEEE Trans. Electron Devices*, 49:1227–1231, 2002.

[49] K. Romanjek, F. Andrieu, T. Ernst, and G. Ghibaudo. Improved split $C-V$ method or effective mobility extraction in sub-0.1 μm Si MOSFETs. *IEEE Electron Device Lett.*, 25:583585, 2004.

[50] M. S. Lundstrom. Elementary scattering theory of the Si MOSFET. *IEEE Electron Device Lett.*, pages 361–363, 1997.

[51] *MASTAR 4: Model for Assessment of cmoS Technologies And Roadmaps, available at:* `http://www.itrs.net/models.html`.

[52] Y. Taur and T. H. Ning. *Fundamentals of Modern VLSI Devices.* Cambridge University Press, 1998.

[53] M. Khare, S.H. andKu, R.A. Donaton, S. Greco, C. Brodsky, X. Chen, A. Chou, R. DellaGuardia, S. Deshpande, B. Doris, S.K.H. Fung, A. Gabor, M. Gribelyuk, S. Holmes, F.F. Jamin, W.L. Lai, W.H. Lee, Y. Li, P. McFarland, R. Mo, S. Mittl, S. Narasimha, D. Nielsen, R. Purtell, W. Rausch, S. Sankaran, J. Snare, L. Tsou, A. Vayshenker, T. Wagner, D. Wehella-Gamage, E. Wu, S. Wu, W. Yan, E. Barth, R. Ferguson, P. Gilbert, D. Schepis, A. Sekiguchi, R. Goldblatt, J. Welser, K.P. Muller, and P Agnello. A high performance 90nm SOI technology with 0.992 μm^2 6T-SRAM cell. In *IEDM Tech. Dig.*, pages 407–410, 2002.

[54] R.A. Chapman, J.W. Kuehne, P.S.-H. Ying, W.F. Richardson, A.R. Paterson, A.P. Lane, I.-C. Chen, L. Velo, C.H. Blanton, M.M. Mosiehl, and J.L. Paterson. High performance sub-half micron CMOS using rapid thermal processing. In *IEDM Tech. Dig.*, pages 101–104, 1991.

[55] Y. Taur, S. Wind, Y.J. Mii, Y. Lii, D. Moy, K.A. Jenkins, C.L. Chen, P.J. Coane, D. Klaus, J. Bucchignano, M. Rosenfield, M.G.R. Thomson, and M. Pol-

cari. High performance 0.1 μm CMOS devices with 1.5 V power supply. In *IEDM Tech. Dig.*, pages 127–130, 1993.

[56] M.S.C. Luo, P.V.G. Tsui, W.-M. Chen, P.V. Gilbert, B. Maiti, A.R. Sitaram, and S.-W. Sun. A 0.25 μm cmos technology with 45 Å NO-nitrided oxide. In *IEDM Tech. Dig.*, pages 691–694, 1995.

[57] M. Rodder, Q.Z. Hong, M. Nandakumar, S. Aur, J.C. Hu, and I.-C. Chen. A sub-0.18μm gate length CMOS technology for high performance (1.5V) and low power (1.0V). In *IEDM Tech. Dig.*, pages 563–566, 1996.

[58] L. Su, R. Schulz, J. Adkisson, K. Beyer, G. Biery, W. Cote, E. Crabbe, D. Edelstein, J. Ellis-Monaghan, E. Eld, D. Foster, R. Gehres, R. Goldblatt, N. Greco, C. Guenther, J. Heidenreich, J. Herman, D. Kiesling, L. Lin, S.-H. Lo, J. McKenna, C. Megivern, H. Ng, J. Oberschmidt, A. Ray, N. Rohrer, K. Tallman, T. Wagner, and B. Davari. A high-performance sub-0.25 μm cmos technology with multiple thresholds and copper interconnects. In *Symp. VLSI Tech.*, pages 18–19, 1998.

[59] M. Hargrove, S. Crowder, E. Nowak, R. Logan, L.K. Han, H. Ng, A. Ray, D. Sinitsky, P. Smeys, F. Guarin, J. Oberschmidt, E. Crabbe, D. Yee, and L. Su. High-performance sub-0.08 μm CMOS with dual gate oxide and 9.7 ps inverter delay. In *IEDM Tech. Dig.*, pages 627–630, 1998.

[60] P. Gilbert, I. Yang, C. Pettinato, M. Angyal, B. Boeck, C. Fu, T. VanGompel, R. Tiwari, T. Sparks, W. Clark, C. Dang, J. Mendonca, B. Chu, K. Lucas, M. Kling, B. Roman, E. Park, F. Huang, M. Woods, D. Rose, K. McGuffin, A. Nghiem, E. Banks, T. McNelly, C. Feng, J. Sturtevant, H. De, A. Das, S. Veeraraghavan, F. Nkansah, and M. Bhat. A high performance l.5V, 0.10μm gate length CMOS technology with scaled copper metalization. In *IEDM Tech. Dig.*, pages 1013–1016, 1998.

[61] E. Leobandung, E. Barth, M. Sherony, S.-H. Lo, R. Schulz, W. Chu, M. Khare, D. Sadana, D. Schepis, R. Boiam, I. Sleight, F. White, F. Assaderaghi, D. Moy,

G. Biery, R. Goldblan, T.-C. Chen, B. Davari, and G. Shahidi. High performance 0.18 μm SOI CMOS technology. In *IEDM Tech. Dig.*, pages 679–682, 1999.

[62] K.K. Young, S.Y. Wu, C.C. Wu, C.H. Wang, C.T. Lin, J.Y. Cheng, M. Chiang, S.H. Chen, T.C. Lo, Y.S. Chen, J.H. Chen, L.J. Chen, S.Y. Hou, J.J. Law, T.E. Chang, C.S. Hou, J. Shih, S.M. Jeng, H.C. Hsieh, Y. Ku, T. Yen, H. Tao, L.C. Chao, S. Shue, S.M. Jang, T.C. Ong, C.H. Yu, M.S. Liang, C.H. Diaz, and J.Y.C. Sun. A 0.13 μm CMOS technology with 193 nm lithography and Cu/low-k for high performance applications. In *IEDM Tech. Dig.*, pages 563–566, 2000.

[63] K. Ichinose, T. Saito, Y. Yanagida, Y. Nonaka, K. Torii, H. Sato, N. Saito, S. Wada, K. Mori, and S. Mitani. A high performance 0.12 μm CMOS with manufacturable 0.18 μm technology. In *Symp. VLSI Tech.*, pages 103–104, 2001.

[64] M. Celik, S. Krishnan, M. Fuselier, A. Wei, D. Wu, B. En, N. Cave, P. Abramowitz, Byoung Min, M. Pelella, Ping Yeh, G. Burbach, B. Taylor, Yongjoo Jeon, Wen-Jie Qi, Ruigang Li, J. Conner, G. Yeap, M. Woo, M. Mendicino, O. Karlsson, and D. Wristers. A 45 nm gate length high performance SOI transistor for 100nm CMOS technology applications. In *Symp. VLSI Tech.*, pages 166–167, 2002.

[65] V. Chan, R. Rengarajan, N. Rovedo, W. Jin, T. Hook, P. Nguyen, J. Chen, E. Nowak, X.-D. Chen, D. Lea, A. Chakravarti, V. Ku, S. Yang, A. Steegen, C. Baiocco, P. Shafer, Ng. H., S.-F. Huang, and C. Wann. High speed 45nm gate length CMOSFETs integrated into a 90nm bulk technology incorporating strain engineering. In *IEDM Tech. Dig.*, pages 77–80, 2003.

[66] K. Mistry, M. Armstrong, C. Auth, S. Cea, T. Coan, T. Ghani, T. Hoffmann, A. Murthy, R. Sandford, J. Shaheed, K. Zawadzki, K. Zhang, S. Thompson, and M. Bohr. Delaying forever: Uniaxial strained silicon transistors in a 90nm CMOS technology. In *Symp. VLSI Tech.*, pages 50–51, 2004.

[67] H.S. Yang, R. Malik, S. Narasimha, Y. Li, R. Divakaruni, P. Agnello, S. Allen, A. Antreasyan, J.C. Arnold, K. Bandy, M. Belyansky, A. Bonnoit, G. Bronner, V. Chan, X. Chen, Z. Chen, D. Chidambarrao, A. Chou, W. Clark, S.W. Crowder, B. Engel, H. Harifuchi, S.F. Huang, R. Jagannathan, F.F. Jamin, Y. Kohyama, H. Kuroda, C.W. Lai, H.K. Lee, W.-H. Lee, E.H. Lim, W. Lai, A. Mallikarjunan, K. Matsumoto, A. McKnight, J. Nayak, H.Y. Ng, S. Panda, R. Rengarajan, M. Steigerwalt, S. Subbanna, K. Subramanian, J. Sudijono, G. Sudo, S.-P. Sun, B. Tessier, Y. Toyoshima, P. Tran, R. Wise, R. Wong, I.Y. Yang, C.H. Wann, L.T. Su, M. Horstmann, Th. Feudel, A. Wei, K. Frohberg, G. Burbach, M. Gerhardt, M. Lenski, R. Stephan, K. Wieczorek, M. Schaller, H. Salz, J. Hohage, H. Ruelke, J. Klais, P. Huebler, S. Luning, R. van Bentum, G. Grasshoff, C. Schwan, E. Ehrichs, S. Goad, J. Buller, S. Krishnan, D. Greenlaw, M. Raab, and N. Kepler. Dual stress liner for high performance sub-45nm gate length soi cmos manufacturing. In *IEDM Tech. Dig.*, pages 1075–1077, 2004.

[68] A. Pouydebasque, B. Dumont, S. Denorme, F. Wacquant, M. Bidaud, C. Laviron, A. Halimaoui, C. Chaton, J.D. Chapon, P. Gouraud, F. Leverd, H. Bernard, S. Warrick, D. Delille, K. Romanjek, R. Gwoziecki, N. Planes, S. Vadot, I. Pouilloux, F. Arnaud, F. Boeuf, and T. Skotnicki. High density and high speed SRAM bit-cells and ring oscillators due to laser annealing for 45nm bulk CMOS. In *IEDM Tech. Dig.*, pages 663–666, 2005.

[69] W.-H. Lee, A. Waite, H. Nii, H.M. Nayfeh, V. McGahay, H. Nakayama, D. Fried, H. Chen, L. Black, R. Bolam, J. Cheng, D. Chidambarrao, C. Christiansen, M. Cullinan-Scholl, D.R. Davies, A. Domenicucci, P. Fisher, J. Fitzsimmons, J. Gill, M. Gribelyuk, D. Harmon, J. Holt, K. Ida, M. Kiene, J. Kluth, C. Labelle, A. Madan, K. Malone, P.V. McLaughlin, M. Minami, D. Mocuta, R. Murphy, C. Muzzy, M. Newport, S. Panda, I. Peidous, A. Sakamoto, T. Sato, G. Sudo, H. VanMeer, T. Yamashita, H. Zhu, P. Agnello, G. Bronner, G. Freeman, S.-F. Huang, T. Ivers, S. Luning, K. Miyamoto, H. Nye, J. Pel-

lerin, K. Rim, D. Schepis, T. Spooner, X. Chen, M. Khare, M. Horstmann, A. Wei, T. Kammler, J. Hontschel, H. Bierstedt, H.-J. Engelmann, A. Hellmich, K. Hempel, G. Koerner, A. Neu, R. Otterbach, C. Reichel, M. Trentsch, P. Press, K. Frohberg, M. Schaller, H. Salz, J. Hohage, H. Ruelke, J. Klais, M. Raab, D. Greenlaw, and N. Kepler. High performance 65 nm SOI technology with enhanced transistor strain and advanced-low-K BEOL. In *IEDM Tech. Dig.*, pages 56–59, 2005.

[70] R.A. Chapman, C.C. Wei, D.A. Bell, S. Aur, G.A. Brown, and R.A. Haken. 0.5 micron CMOS for high performance at 3.3 V. In *IEDM Tech. Dig.*, pages 52–55, 1988.

[71] J. Hayden, F. Baker, S. Ernst, B. Jones, J. Klein, M. Lien, T. McNelly, T. Mele, H. Mendez, B.Y. Nguyen, L. Parrillo, W. Paulson, J. Pfiester, F. Pintchovski, Y. See, R. Sivan, B. Somero, and E. Travis. A high-performance sub-half micron CMOS technology for fast SRAMs. In *IEDM Tech. Dig.*, pages 417–420, 1989.

[72] M. Rodder, S. Iyer, S. Aur, A. Chatterjee, J. McKee, R. Chapman, and I.-C. Chen. Oxide thickness dependence of inverter delay and device reliability for 0.25μm CMOS technology. In *IEDM Tech. Dig.*, pages 879–882, 1993.

[73] M. Rodder, A. Amerasekera, S. Aur, and I.-C. Chen. A study of design/process dependence of 0.25μm gate length CMOS for improved performance and reliability. In *IEDM Tech. Dig.*, pages 71–74, 1994.

[74] M. Rodder, S. Aur, and I.-C. Chen. A scaled 1.8 V, 0.18/μm gate length cmos technology: Device design and reliability considerations. In *IEDM Tech. Dig.*, pages 415–418, 1995.

[75] M. Rodder, M. Hanratty, D. Rogers, T. Laaksonen, J.C. Hu, S. Murtaza, C.-P. Chao, S. Hattangady, S. Aur, A. Amerasekera, and I.-C. Chen. A 0.10 μm gate length CMOS technology with 30 å gate dielectric for 1.0 V - 1.5 V applications. In *IEDM Tech. Dig.*, pages 223–226, 1997.

[76] M. Rodder, S. Hattangady, N. Yu, W. Shiau, P. Nicollian, T. Laaksonen, C.P. Chao, M. Mehrotra, C. Lee, S. Murtaza, and S. Aur. A 1.2 V, 0.1 μm gate length CMOS technology: Design and process issues. In *IEDM Tech. Dig.*, pages 623–626, 1998.

[77] M. Mehrotra, J.C. Hu, A. Jain, W. Shiau, V. Reddy, S. Aur, and M Rodder. A 1.2V, sub-0.09 μm gate length CMOS technology. In *IEDM Tech. Dig.*, pages 419–422, 1999.

[78] A.H. Perera, B. Smith, N. Cave, M. Sureddin, S. Chheda, R. Islam, J. Chang, S.-C. Song, A. Sultan, S. Crown, V. Kolagunta, S. Shah, M. Celik, D. Wu, K.C. Yu, R. Fox, S. Park, C. Simpson, D. Eades, S. Gonzales, C. Thomas, J. Sturtevant, D. Bonser, N. Benavides, M. Thompson, V. Sheth, J. Fretwell, S. Kim, N. Ramani, K. Green, M. Moosa, P. Besser, Y. Solomentsev, D. Denning, M. Friedemann, B. Baker, R. Chowdhury, S. Ufmani, K. Strozewski, R. Carter, J. Reiss, M. Olivares, B. Ho, T. Lii, T. Sparks, T. Stephens, M. Schaller, C. Goldberg, K. Junker, D. Wristers, J. Alvis, B. Melnick, and S. Venkatesan. A versatile 0.13 μm CMOS platform technology supporting high performance and low power applications. In *IEDM Tech. Dig.*, pages 571–574, 2000.

[79] N. Yanagiya, S. Matsuda, S. Inaba, M. Takayanagi, I. Mizushima, K. Ohuchi, K. Okano, K. Takahasi, E. Morifuji, M. Kanda, Y. Matsubara, M. Habu, M. Nishigoori, K. Honda, H. Tsuno, K. Yasumoto, T. Yamamoto, K. Hiyama, K. Kokubun, T. Suzuki, J. Yoshikawa, T. Sakurai, T. Ishizuka, Y. Shoda, M. Moriuchi, M. Kishida, H. Matsumori, H. Harakawa, H. Oyamatsu, N. Nagashima, S. Yamada, T. Noguchi, H. Okamoto, and M. Kakumu. 65 nm CMOS technology (CMOS5) with high density embedded memories for broadband. In *IEDM Tech. Dig.*, pages 57–60, 2002.

[80] K. Goto, Y. Tagawa, H. Ohta, H. Morioka, S. Pidin, Y. Momiyama, H. Kokura, S. Inagaki, N. Tamura, M. Hori, T. Mori, M. Kase, K. Hashimoto, M. Kojima,

and T. Sugii. Technology booster using strain-enhancing laminated SiN (SELS) for 65nm node HP MPUs. In *IEDM Tech. Dig.*, pages 623–626, 2003.

[81] E. Leobandung, H. Nayakama, D. Mocuta, K. Miyamoto, M. Angyal, H.V. Meer, K. McStay, I. Ahsan, S. Allen, A. Azuma, M. Belyansky, R.-V. Bentum, J. Cheng, D. Chidambarrao, B. Dirahoui, M. Fukasawa, M. Gerhardt, M. Gribelyuk, S. Halle, H. Harifuchi, D. Harmon, J. Heaps-Nelson, H. Hichri, K. Ida, M. Inohara, I.C. Inouc, K. Jenkins, T. Kawamura, B. Kim, S.-K. Ku, M. Kumar, S. Lane, L. Liebmann, R. Logan, I. Melville, K. Miyashita, A. Mocuta, P. Oapos andNeil, M.-F. Ng, T. Nogami, A. Nomura, C. Norris, E. Nowak, M. Ono, S. Panda, C. Penny, C. Radens, R. Ramachandran, A. Ray, S.-H. Rhee, D. Ryan, T. Shinohara, G. Sudo, F. Sugaya, J. Strane, Y. Tan, L. Tsou, L. Wang, F. Wirbeleit, S. Wu, T. Yamashita, H. Yan, Q. Ye, D. Yoneyama, D. Zamdmer, H. Zhong, H. Zhu, W. Zhu, P. Agnello, S. Bukofsky, G. Bronner, E. Crabbe, G. Freeman, S.-F. Huang, T. Ivers, H. Kuroda, D. McHerron, J. Pellerin, Y. Toyoshima, S. Subbanna, N. Kepler, and L. Su. High performance 65 nm SOI technology with dual stress liner and low capacitance SRAM cell. In *Symp. VLSI Tech.*, pages 126–127, 2005.

[82] M. Okuno, K. Okabe, T. Sakuma, K. Suzuki, T. Miyashita, T. Yao, H. Morioka, M. Terahara, Y. Kojima, H. Watatani, K. Sugimoto, T. Watanabe, Y. Hayami, T. Mori, T. Kubo, Y. Iba, I. Sugiura, H. Fukutome, Y. Morisaki, H. Minakata, K. Ikeda, S. Kishii, N. Shimizu, T. Tanaka, S. Asai, M. Nakaishi, S. Fukuyama, A. Tsukune, M. Yamabe, I. Hanyuu, M. Miyajima, M. Kase, K. Watanabe, S. Satoh, and T. Sugii. 45-nm cmos integration with a novel STI structure and full-NCS/Cu interlayers for low-operation-power (LOP) applications. In *IEDM Tech. Dig.*, pages 52–55, 2005.

[83] K. Adachi, K. Ohuchi, N. Aoki, H. Tsujii, T. Ito, H. Itokawa, K. Matsuo, K. Suguro, Y. Honguh, N. Tamaoki, K. Ishimaru, and H. Ishiuchi. Issues and optimization of millisecond anneal process for 45 nm node and beyond. In *Symp. VLSI Tech.*, pages 142–143, 2005.

[84] S. Yu, J.-P. Lu, F. Mehrad, H. Bu, A. Shanware, M. Ramin, M. Pas, M.R. Visokay, S. Vitale, S.-H. Yang, P. Jiang, L. Hall, C. Montgomery, Y. Obeng, C. Bowen, H. Hong, J. Tran, R. Chapman, S. Bushman, C. Machala, J. Blatchford, R. Kraft, L. Colombo, S. Johnson, and B. McKee. 45-nm node NiSi FUSI on nitrided oxide bulk CMOS fabricated by a novel integration process. In *IEDM Tech. Dig.*, pages 231–234, 2005.

[85] P. Ranade, T. Ghani, K. Kuhn, K. Mistry, S. Pae, L. Shifren, M. Stettler, K. Tone, S. Tyagi, and M. Bohr. High performance 35nm L_{GATE} CMOS transistors featuring NiSi metal gate (FUSI), uniaxial strained silicon channels and 1.2nm gate oxide. In *IEDM Tech. Dig.*, pages 227–230, 2005.

[86] M. Horstmann, A. Wei, T. Kammler, J. Hntschel, H. Bierstedt, T. Feudel, K. Frohberg, M. Gerhardt, A. Hellmich, K. Hempel, J. Hohage, P. Javorka, J. Klais, G. Koerner, M. Lenski, A. Neu, R. Otterbach, P. Press, C. Reichel, M. Trentsch, B. Trui, H. Salz, M. Schaller, H.-J. Engelmann, O. Herzog, H. Ruelke, P. Hubler, R. Stephan, D. Greenlaw, M. Raab, and N. Kepler. Integration and optimization of embedded-SiGe, compressive and tensile stressed liner films, and stress memorization in advanced SOI CMOS technologies. In *IEDM Tech. Dig.*, pages 243–246, 2005.

[87] H. Ohta, Y. Kim, Y. Shimamune, T. Sakuma, A. Hatada, A. Katakami, T. Soeda, K. Kawamura, H. Kokura, H. Morioka, T. Watanabe, J.O.Y. Hayami, J. Ogura, M. Tajima, T. Mori, N. Tamura, M. Kojima, and K. Hashimoto. High performance 30 nm gate bulk CMOS for 45 nm node with Σ-shaped SiGe S/D. In *IEDM Tech. Dig.*, pages 247–250, 2005.

[88] T. Kinoshita, R. Hasumi, M. Hamaguchi, K. Miyashita, T. Komoda, A. Kinoshita, J. Koga, K. Adachi, Y. Toyoshima, T. Nakayama, S. Yamada, and F. Matsuoka. Ultra low voltage operations in bulk cmos logic circuits with dopant segregated schottky source/drain transistors. In *IEDM Tech. Dig.*, pages 71–74, 2006.

[89] S. Narasimha, K. Onishi, H. M. Nayfeh, A. Waite, M. Weybright, J. Johnson, C. Fonseca, D. Corliss, C. Robinson, M. Crouse, D. Yang, C-H.J. Wu, A. Gabor, T. Adam, I. Ahsan, M. Belyansky, L. Black, S. Butt, J. Cheng, A. Chou, G. Costrini, C. Dimitrakopoulos, A. Domenicucci, P. Fisher, A. Frye, S. Gates, S. Greco, S. Grunow, M. Hargrove, J. Holt, S-J. Jeng, M. Kelling, B. Kim, W. Landers, G. Larosa, D. Lea, M.H. Lee, X. Liu, N. Lustig, A. McKnight, L. Nicholson, D. Nielsen, K. Nummy, V. Ontalus, C. Ouyang, X. Ouyang, C. Prindle, R. Pal, W. Rausch, D. Restaino, C. Sheraw, J. Sim, A. Simon, T. Standaert, C. Y. Sung, K. Tabakman, C. Tian, R. Van Den Nieuwenhuizen, H. Van Meer, A. Vayshenker, D. Wehella-Gamage, J. Werking, R. C. Wong, S. Wu, J. Yu, R. Augur, D. Brown, X. Chen, D. Edelstein, A. Grill, M. Khare, Y. Li, S. Luning, J. Norum, S. Sankaran, D. Schepis, R. Wachnik, R. Wise, C. Wann, T. Ivers, and P. Agnello. High performance 45-nm SOI technology with enhanced strain, porous low-k BEOL, and immersion lithography. In *IEDM Tech. Dig.*, pages 689–692, 2006.

[90] P. Palestri, D. Esseni, S. Eminente, C. Fiegna, E. Sangiorgi, and L. Selmi. Understanding quasi-ballistic transport in nano-MOSFETs: part I-scattering in the channel and in the drain. *IEEE Trans. Electron Devices*, 52(12):2727–2735, 2005.

[91] P. Palestri, R. Clerc, D. Esseni, L. Lucci, and L. Selmi. Multi-subband-monte-carlo investigation of the mean free path and of the kT layer in degenerated quasi ballistic nanomosfets. In *IEDM Tech. Dig.*, pages 933–936, 2006.

[92] J.-P. Han, H. Utomo, L. W. Teo, N. Rovedo, Z. Luo, R. Krishnasamy, R. Stierstorfer, Y. F. Chong, S. Fang, H. Ng, J. Holt, T. N. Adam, J. Kempisty, A. Gutmann, D. Schepis, S. Mishra, H. Zhuang, J. J. Kim, J. Li, R. Murphy, R. Davis, B. St. Lawrence, A. Madan, A. Turansky, L. Burns, R. Loesing, S.-D. Kim, R. Lindsay, G. Chiulli, R. Amos, M. Hierlemann, D. Shum, J. H. Ku, J. Sudijono, and M. Ieong. Novel enhanced stressor with graded embedded

sige source/drain for high performance CMOS devices. In *IEDM Tech. Dig.*, pages 59–62, 2006.

[93] Y. Tateshita, J. Wang, K. Nagano, T. Hirano, Y. Miyanami, T. Ikuta, T. Kataoka, Y. Kikuchi, S. Yamaguchi, T. Ando, K. Tai, R. Matsumoto, S. Fujita, C. Yamane, R. Yamamoto, S. Kanda, K. Kugimiya, T. Kimura, T. Ohchi, Y. Yamamoto, Y. Nagahama, Y. Hagimoto, H. Wakabayashi, Y. Tagawa, M. Tsukamoto, H. Iwamoto, M. Saito, S. Kadomura, and N. Nagashima. High-performance and low-power CMOS device technologies featuring metal/high-k gate stacks with uniaxial strained channels on (100) and (110) substrates. In *IEDM Tech. Dig.*, pages 63–66, 2006.

[94] H. C.-H. Wang, S.-H. Huang, C.-W. Tsai, H.-H. Lin, T.-L. Lee, S.-C. Chen, C. H. Diaz, M.-S. Liang, and J. Y.-C. Sun. High performance PMOS devices on (110)/$\langle 111' \rangle$ substrate/channel with multiple stressors. In *IEDM Tech. Dig.*, pages 67–70, 2006.

[95] H. Hu, J. Jacobs, L. Su, and D. A. Antoniadis. A study of deep-submicron MOSFET scaling based on experiment and simulation. *IEEE Trans. Electron Devices*, 42:669–677, 1995.

[96] J. W. Sleight, I. Lauer, O. Dokumaci, D. M. Fried, D. Guo, B. Haran, S. Narasimha, C. Sheraw, D. Singh, M. Steigerwalt, X. Wang, P. Oldiges, D. Sadana, C. Y. Sung, W. Haensch, and M. Khare. Challenges and opportunities for high performance 32 nm CMOS technology, 2006.

[97] N. R. Mohapatra, M. P. Desai, S. G. Narendra, and V. R. Rao. Modeling of parasitic capacitance in deep submicrometer conventional and high-k dielectric MOS transistors. *IEEE Trans. Electron Devices*, 50:959–966, 2003.

[98] M. Togo, A. Tanabe, A. Furukawa, K. Tokunaga, and T. Hashimoto. A gate-side air-gap structure (GAS) to reduce the parasitic capacitance in MOSFETs. In *VLSI Tech. Symp.*, pages 38–90, 1998.

[99] S. L. Wu. MOSFET with self-aligned silicidation and gate-side air-gap structure, US Patent 5,915,182.

[100] S.-D. Kim, C.-M. Park, and J.C.S. Woo. Advanced model and analysis of series resistance for CMOS scaling into nanometer regime — Part I. theoretical derivation. *IEEE Trans. Electron Devices*, 49(3):457–466, 2002.

[101] M. S. Lundstrom and Ren. Z. Essential physics of carrier transport in nanoscale MOSFET's. *IEEE Trans. Electron Devices*, pages 133–141, 2002.

[102] A. Svizhenko and M. P. Anantram. Role of scattering in nanotransistors. *IEEE Trans. Electron Devices*, 50:1459– 1466, 2003.

[103] F. Assad, Z. Ren, D. Vasileska, S. Datta, and M. Lundstrom. On the performance limits for silicon MOSFETs: A theoretical study. *IEEE Trans. Electron Devices*, 47:232–240, 2000.

[104] A. Rahman and M. S. Lundstrom. A compact scattering model for the nanoscale double-gate MOSFET. *IEEE Trans. Electron Devices*, 49(3):481–489, 2002.

[105] M.S. Lundstrom. On the mobility versus drain current relation for a nanoscale MOSFET. *IEEE Electron Device Lett.*, 22(6):293–295, 2001.

[106] A. Lochtefeld and D. A. Antoniadis. Investigating the relationship between electron mobility and velocity in deeply scaled NMOS via mechanical stress. *IEEE Electron Device Lett.*, 22(12):591–593, 2001.

[107] S. Takagi, J. L. Hoyt, J. J. Welser, and J. F. Gibbons. Comparative study of phonon-limited mobility of two-dimensional electrons in strained and unstrained si metaloxidesemiconductor field-effect transistors. *J. Appl. Phys.*, 80(3):1567–1577, 1994.

[108] T. Kobayashi and K. Saito. Two-dimensional analysis of velocity overshoot effects in ultrashort-channel Si MOSFET's. *IEEE Trans. Electron Devices*, 32:788– 792, 1985.

[109] Z. Ren, R. Venugopal, S. Goasguen, S. Datta, and M. S. Lundstrom. nanoMOS 2.5: A two-dimensional simulator for quantum transport in double-gate MOSFETs. *IEEE Trans. Electron Devices*, 50:1914–1925, 2003.

[110] K. Uchida, T. Krishnamohan, K.C. Saraswat, and Y. Nishi. Physical mechanisms of electron mobility enhancement in uniaxial stressed MOSFETs and impact of uniaxial stress engineering in ballistic regime. In *IEDM Tech. Dig.*, pages 129–132, 2005.

[111] F. Rochette, M. Casse, M. Mouis, D. Blachier, C. Leroux, B. Guillaumot, G. Reimbold, and F. Boulanger. Electron mobility enhancement in uniaxially strained MOSFETs: Extraction of the effective mass variation. In *Proc. European Solid-State Device Research Conf.*, pages 93–96, 2006.

[112] S. Takagi and A. Toriumi. Quantitative understanding of inversion-layer capacitance in Si MOSFET's. *IEEE Trans. Electron Devices*, 42(12):2125–2130, 1995.

[113] F. Rana, S. Tiwari, and D. A. Buchanan. Self-consistent modeling of accumulation layers and tunneling currents through very thin oxides. *Appl. Phys. Lett.*, 69(8):1104–1106, 1996.

[114] X. Yang, J. Lim, G. Sun, K. Wu, T. Nishida, and S. E. Thompson. Strain-induced changes in the gate tunneling currents in p-channel metaloxidesemiconductor field-effect transistors. *Appl. Phys. Lett.*, 88:052108, 2006.

[115] D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H.-S.P. Wong. Device scaling limits of Si MOSFETs and their application dependencies. *IEEE Proc.*, 89(3):259–288, 2001.

[116] A. Rahman, A. Ghosh, and M. Lundstrom. Assessment of Ge n-MOSFETs by quantum simulation. In *IEDM Tech. Dig.*, pages 471–474, 2003.

[117] T. Krishnamohan, D. Kim, C. D. Nguyen, C. Jungemann, Y. Nishi, and K.C. Saraswat. High-mobility low band-to-band-tunneling strained-germanium

double-gate heterostructure FETs: Simulations. *IEEE Trans. Electron devices*, 53:1000–1009, 2006.

[118] D. Vasileska and Z. Ren. *Schred 2.0 Manual: A 1D, Self-Consistent Schrdinger-Poisson Solver Including Many Body and Exchange Effects for bulk and SOI structures.* Arizona State University, Purdue University, 2000.

[119] A. Khakifirooz and D. A. Antoniadis. On the electron mobility in ultrathin SOI and GOI. *IEEE Electron Device Lett.*, 25:80–82, 2004.

[120] K. Uchida, H. Watanabe, A. Kinoshita, J. Koga, T. Numata, and S. Takagi. Experimental study on carrier transport mechanism in ultrathin-body SOI n and p-MOSFETs with SOI thickness less than 5 nm. In *IEDM Tech. Dig.*, pages 47–50, 2002.

[121] K. Rim, J. Chu, H. Chen, K.A. Jenkins, T. Kanarsky, K. Lee, A. Mocuta, H. Zhu, R. Roy, J. Newbury, J. Ott, K. Petrarca, P. Mooney, D. Lacey, S. Koester, K. Chan, D. Boyd, M. Ieong, and H.-S. Wong. Characteristics and device design of sub-100 nm strained si N- and PMOSFETs. In *Symp. VLSI Tech.*, pages 98–99, 2002.

[122] J.R. Hwang, J.H. Ho, S.M. Ting, T.P. Chen, Y.S. Hsieh, C.C. Huang, Y.Y. Chiang, H.K. Lee, Ariel Liu, T.M. Shen, G. Braithwaite, M. Currie, N. Gerrish, R. Hammond, A. Lochtefeld, F. Singaporewala, M. Bulsara, Q. Xiang, M.R. Lin, W.T. Shiau, Y.T. Loh, J.K. Chen, and S.C. Chien. Performance of 70 nm strained-silicon CMOS devices. In *Symp. VLSI Tech.*, pages 103–104, 2003.

[123] M. V. Fischetti, F. Gàmiz, and W. Haensch. On the enhanced electron mobility in strained-silicon inversion layers. *J. Applied Physics*, 92(12):7320–7324, 2002.

[124] O. Bonno, S. Barraud, F. Andrieu, D. Mariolle, F. Rochette, M. Casse, J. M. Hartmann, F. Bertin, and O. Faynot. High-field electron mobility in biaxially-tensile strained SOI: low temperature measurement and correlation with surface morphology. In *Symp. VLSI Tech.*, pages 134–135, 2007.

[125] H.M. Nayfeh, C.W. Leitz, A.J. Pitera, E.A. Fitzgerald, J.L. Hoyt, and D.A. Antoniadis. Influence of high channel doping on the inversion layer electron mobility in strained silicon n-MOSFETs. *IEEE Electron Device Lett.*, 24(4):248–250, 2003.

[126] O. Weber and S. Takagi. New findings on coulomb scattering mobility in strained-Si nFETS and its physical understandu. In *Symp. VLSI Tech.*, pages 130–131, 2007.

[127] I. Lauer and D. A. Antoniadis. Enhancement of electron mobility in ultrathin-body silicon-on-insulator MOSFETs with uniaxial strain. *IEEE Electron Device Lett.*, 26:314–316, 2005.

[128] H. Yin, Z. Ren, H. Chen, J. Holt, X. Liu, J.W. Sleight, K. Rim, V. Chan, D.M. Fried, Y.H. Kim, J.O. Chu, B.J. Greene, S.W. Bedell, G. Pfeiffer, R. Bendernagel, D.K. Sadana, T. Kanarsky, C.Y. Sung, M. Ieong, and G. Shahidi. Integration of local stress techniques with strained-Si directly on insulator (SSDOI) substrates. In *VLSI Symp. Tech.*, pages 76–77, 2006.

[129] A.V.-Y. Thean, L. Prabhu, V. Vartanian, M. Ramon, B.-Y. Nguyen, T. White, H. Collard, Q.-H. Xie, S. Murphy, J. Cheek, S. Venkatesan, J. Mogab, C.H. Chang, Y.H. Chiu, H.C. Tuan, Y.C. See, M.S. Liang, and Y.C. Sun. Uniaxial-biaxial stress hybridization for super-critical strained-Si directly on insulator (SC-SSOI) PMOS with different channel orientations. In *IEDM Tech. Dig.*, pages 509–512, 2005.

[130] K. Ota, K. Sugihara, H. Sayama, T. Uchida, H. Oda, T. Eimori, H. Morimoto, and Y. Inoue. Novel locally strained channel technique for high performance 55nm CMOS. In *IEDM Tech. Dig.*, pages 27–30, 2002.

[131] E. Ungersboeck, S. Dhar, G. Karlowatz, H. Kosina, and S. Selberherr. Physical modeling of electron mobility enhancement for arbitrarily strained silicon. *J. Computational Electronics*, 6(1-3):55–58, 2007.

[132] T. B. Boykin, G. Klimeck, and F. Oyafuso. Valence band effective mass expressions in the $sp^3d^5s^*$ empirical tight-binding model applied to a new Si and Ge parameterization. *Phys. Rev. B*, 69:115201, 2004.

[133] D. Rideau, M. Feraille, L. Ciampolini, M. Minondo, C. Tavernier, H. Jaouen, and A. Ghetti. Strained Si, Ge, and $Si_{1-x}Ge_x$ alloys modeled with a first-principles-optimized full-zone kp method. *Phys. Rev. B*, 74:195208, 2006.

[134] I. Lauer. *The effects of strain on carrier transport in thin and ultra-thin SOI MOSFETs.* PhD thesis, Massachusetts Institute of Technology, 2005.

[135] J. J. Rosenberg and S. C. Martin. Self-aligned germanium MOSFET's using a nitrided native oxide gate insulator. *IEEE Electron Device Lett.*, 9(12):639–641, 1988.

[136] S. C. Martin, L. M. Hitt, and J. J. Rosenberg. p-channel germanium MOSFET's with high channel mobility. *IEEE Electron Device Lett.*, 10(7):325–327, 1989.

[137] T. N. Jackson, C. M. Ransom, and J. F. DeGelormo. Gate-self-aligned p-channel germanium MISFET's. *IEEE Electron Device Lett.*, 12(11):605–607, 1991.

[138] A. Ritenour. *Design, Fabrication, and Characterization of Germanium MOSFETs with High-k Gate Dielectric Stacks based on Nitride Interfacial Layers.* PhD thesis, Massachusetts Institute of Technology, 2007.

[139] K. Fletcher and G. D. Pitt. Intervalley scattering in n type Ge from a Hall effect experiment to high pressures. *J. Phys. C: Solid State Phys.*, 4:1822–1834, 1971.

[140] W. Fawcett and E. G. S. Paige. Negative differential mobility of electrons in germanium: A monte carlo calculation of the distribution function, drift velocity and carrier population in the (111) and (100) minima. *J. Phys. C: Solid State Phys.*, 4:1801–1821, 1971.

[141] M. I. Nathan, W. Paul, and H. Brooks. Interband scattering in n-type germanium. *Phys. Rev.*, 124(2):391–407, 1961.

[142] M. I. Daunov, I. K. Kamilov, S. F. Gabibov, and P. K. Akchurin. Interband electron scattering in germanium with a doubly charged level of gold as judged from a study of hall mobility under high pressure. *Semiconductor Science and Technology*, 17(3):211–214, 2002.

[143] C. N. Ahmad and A. R. Adams. Electron transport and pressure coefficients associated with the L_{1C} and Δ_{1C} minima of germanium. *Phys. Rev. B*, 34(4):2319–2328, 1986.

[144] H. Heinrich, K. Lischka, and M. Kriechbaum. Magnetoresistance and hall effect of hot electrons in germanium and carrier transfer to higher minima. *Phys. Rev. B*, 2:2009 – 2016, 1970.

[145] M. Fischeti and S. E. Laux. Monte Carlo simulation of transport in technological significant semiconductors of the diamond and zinc-blende structures. Part II: Submicrometer MOSFET's. *IEEE Trans. Electron Devices*, 38:650–660, 1991.

[146] C. Jacoboni, F. Nava, C. Canali, and G. Ottaviani. Electron drift velocity and diffusivity in germanium. *Phys. Rev. B*, 24:1014–1026, 1981.

[147] D. Chattopadhyay and B. R. Nag. High-field Hall factor of n-Ge at 200°k. *Phys. Rev. B.*, 4:1220–1223, 1971.

[148] O. Weber, P. Scheiblin, R. Ritzenthaler, T. Ernst, F. Andrieu, F. Ducroquet, J.-F. Damlencourt, Y. Le Tiec, A.-M. Papon, H. Dansas, L. Brevard, A. Toffoli, B. Guillaumot, and S. Deleonibus. A novel locally engineered (111) V-channel pMOSFET architecture with improved drivability characteristics for low-standby power (LSTP) CMOS applications. In *Symp. VLSI Tech.*, pages 156–157, 2005.

[149] S. Takagi, A. Toriumi, M. Iwase, and H. Tango. On the universality of inversion layer mobility in Si MOSFET's: Part II –effects of surface orientation. *IEEE Trans. Electron Devices*, 41(12):2363–, 1994.

[150] M. Uchida, Y. Kamakura, and K. Taniguchi. Performance enhancement of pmosfets depending on strain, channel direction, and material. In *SISPAD*, pages 315–318, 2005.

[151] M. H. Liao, S. T. Chang, M. H. Lee, S. Maikap, and C. W. Liu. Abnormal hole mobility of biaxial strained Si. *J. Appl. Phys.*, 98:066104, 2005.

[152] A. Rahman, G. Klimeck, T. B. Boykin, and M. Lundstrom. Bandstructure effects in ballistic nanoscale MOSFETs. In *IEDM Tech. Dig.*, pages 139–142, 2004.

[153] M. Stadele, A. Di Carlo, P. Lugli, F. Sacconi, and B. Tuttle. Atomistic tight-binding calculations for the investigation of transport in extremely scaled SOI transistors. In *IEDM Tech. Dig.*, pages 229–232, 2003.

[154] I. Åberg and J. L. Hoyt. Hole transport in UTB MOSFETs in strained-Si directly on insulator with strained-Si thickness less than 5 nm. *IEEE Electron Device Lett.*, 26(9):661–663, 2005.

[155] M. V. Fischetti, Z. Ren, P. M. Solomon, M. Yang, and K. Rim. Six-band k.p calculation of the hole mobility in silicon inversion layers: Dependence on surface orientation, strain, and silicon thickness. *J. Appl. Phys.*, 94:1079–1095, 2003.

[156] A. Khakifirooz and D. A. Antoniadis. Scalability of hole mobility enhancement in biaxially strained ultrathin body SOI. *IEEE Electron Device Lett.*, 27:402–404, 2006.

[157] H. Nakatsuji, Y. Kamakura, and K. Taniguchi. A study of subband structure and transport of two-dimensional holes in strained si p-MOSFETs using full-band modeling. In *IEDM Tech. Dig.*, pages 727–730, 2002.

[158] G. A. Garcia, R. E. Reedy, and M. L. Burgener. High-quality CMOS in thin (100 nm) silicon on sapphire. *IEEE Electron Device Lett.*, 9:32–34, 1988.

[159] P. Zimmerman, G. Nicholas, B. De Jaeger, B. Kaczer, A. Stesmans, L.-A. Ragnarsson, D. P. Brunco, F. E. Leys, M. Caymax, G. Winderickx, K. Opsomer, M. Meuris, and M. M. Heyns. High performance Ge pMOS devices using a Si-compatible process flow. In *IEDM Tech. Dig.*, 2006.

[160] A. Rahman, J. Guo, S. Datta, and M. Lundstrom. Theory of ballistic nanotransistors. *IEEE Trans. Electron Devices*, 50:1853–1864, 2003.

[161] F. Andrieu, T. Ernst, F. Lime, F. Rochette, K. Romanjek, S. Barraud, C. Ravit, F. Boeuf, M. Jurczak, M. Casse, O. Weber, L. Brevard, G. Reimbold, G. Ghibaudo, and S. Deleonibus. Experimental and comparative investigation of low and high field transport in substrate- and process-induced strained nanoscaled MOSFETs. In *Symp. VLSI Tech.*, pages 176–177, 2005.

[162] D. V. Singh, J. M. Hergenrother, J. W. Sleight, Z. Ren, H. M. Nayfeh, O. Dokumaci, L. Black, D. Chidambarrao, R. Venigalla, J. Pan, B. L. Tessier, A. Nomura, J. A. Ott, M. Khare., K. W. Guarini, M. Ieong, and W. Haensch. Effect of contact liner stress in high-performance fdsoi devices with ultra-thin silicon channels and 30 nm gate lengths. In *Proc. SOI Conf.*, pages 178–179, 2005.

[163] M. Shima, K. Okabe, A. Yamaguchi, T. Sakoda, K. Kawamura, S. Pidin, M. Okuno, T. Owada, K. Sugimoto, J. Ogura, H. Kokura, H. Morioka, T. Watanabe, T. Isome, K. Okoshi, T. Mori, Y. Hayami, H. Minakata, A. Hatada, Y. Shimamune, A. Katakami, and H. Ota Sakum. High-performance low operation power transistor for 45nm node universal applications. In *Symp. VLSI Tech.*, pages 156–157, 2006.

[164] X. Chen, S. Fang, W. Gao, T. Dyer, Y. Teh, S. Tan, Y. Ko, C. Baiocco, A. Ajmera, J. Park, J. Kim, R. Stierstorfer, D. Chidambarrao, Z. Luo, N. Nivo, P. Nguyen, J. Yuan, S. Panda, O. Kwon, N. Edleman, T. Tjoa, J. Widodo,

M. Belyansky, and M. Sherony. Stress proximity technique for performance improvement with dual stress liner at 45nm technology and beyond. In *Symp. VLSI Tech*, pages 60–61, 2006.

[165] F. Andrieu, T. Ernst, C. Ravit, M. Jurczak, G. Ghibaudo, and S. Deleonibus. In-depth characterization of the hole mobility in 50-nm process-induced strained MOSFETs. *IEEE Electron Device Lett.*, 26:755–757, 2005.

[166] V. Moroz, Z. Krivokapic, X. Xu, D. Pramanik, and F. Nouri. Analyzing strained-silicon options for stress-engineering transistors. *Solid State Technology*, page 49, 2004.

[167] M. V. Fischetti, D. A. Neumayer, and E. A. Cartier. Effective electron mobility in si inversion layers in metaloxidesemiconductor systems with a high-κ insulator: The role of remote phonon scattering. *J. Appl. Phys.*, 90(9):4587–4608, 2001.

[168] M. Hiratani, S. Saito, Y. Shimamoto, and K. Torii. Effective electron mobility reduced by remote charge scattering in high-κ gate stacks. *Jap. J. Appl. Phys.*, 41(7A):4521–4522, 2002.

[169] J. Zhu, J. P. Han, and T. P. Ma. Mobility measurement and degradation mechanisms of MOSFETs made with ultra-thin high-k dielectrics. *IEEE Trans. Electron Devices*, 51(1):98–105, 2004.

[170] B. H. Lee, C. D. Young, R. Choi, J. H. Sim, G. Bersuker, C. K. Kang, R. harris, G. A. Brown, K. Mattews, S. C. Song, N. Moumen, J. Barnett, P. Lysaght, K. S. Choi, H. C. Wen, C. Huffman, H. Alshareef, P. Majhi, S. Gopalan, J. Peterson, P. Kirsh, H.-J. Li, J. Gutt, M. Gardner, H. R. Huff, P. Zeitzoff, R. W. Murto, L. Larson, and C. Ramiller. Intrinsic characteristics of high-k devices and implications of fast transient charging effects (FTCE). In *IEDM Tech. Dig.*, pages 859–862, 2004.

[171] A. Ritenour, J. Hennessy, and D. A. Antoniadis. Investigation of carrier transport in germanium MOSFETs with WN/Al_2O_3/AlN gate stack. *IEEE Electron Device Lett.*, 28:746–749, 2007.

[172] D.V. Singh, P. Solomon, E.P. Gusev, G. Singco, and Z. Ren. Ultra-fast measurements of the inversion charge in MOSFETs and impact on measured mobility in high-k MOSFETs. In *IEDM Tech. Dig.*, pages 863–866, 2004.

[173] G. Groesneken, H. E. Maes, N. Beltra'n, and R. F. D. Keersmaecker. A reliable approach to charge-pumping measurements in MOS transistors. *IEEE Trans. Electron Devices*, 31:42–53, 1984.

[174] A. Kerber, E. Cartier, L. Pantisano, R. Degraeve, T. Kauerauf, Y. Kim, A. Hou, G. Groeseneken, H. E. Maes, and U. Schwalke. Origin of the threshold voltage instability in SiO_2/HFO_2 dual layer gate dielectrics. *IEEE Electron Device Lett.*, 24:87–89, 2003.

[175] D. Heh, C. D. Young, G. A. Brown, P. Y. Hung, A. Diebold, E. M. Vogel, J. B. Bernstein, and G. Bersuker. Spatial distribution of trapping centers in HFO_2/SiO_2 gate stack. *IEEE Trans. Electron Devices*, 54:1338–1345, 2007.

[176] S. Jakschik, A. Avellan, U. Schroeder, and J. W. Bartha. Influesnce of Al_2O_3 dielectrics on the trap-depth profiles in MOS devices investigated by the charge-pumping method. *IEEE Trans. Electron Devices*, 51:2252–2255, 2004.

[177] F. P. Heimann and G. Warfield. The effect of oxide traps on MOS capacitance. *IEEE Trans. Electron Devices*, 12:167–178, 1964.

[178] G. Van den Bosch, G. Groeseneken, P. Heremans, and H. Maes. Spectroscopic charge pumping: a new procedure for measuring interface trap distributions in MOS transistors. *IEEE Trans. Electron Devices*, 38:1820–1831, 1991.

[179] N. R. Mohapatra, M. P. Desai, S. G. Narendra, and V. R. Rao. Modeling of parasitic capacitances in deep submicrometer conventional and high-k dielectric MOS transistors. *IEEE Trans. Electron devices*, 50(4):959 – 966, 2003.

[180] R. G. Humphreys. Valence band averages in silicon: anisotropy and non-parabolicity. *J. Phys. C: Solid State Phys.*, 14:2935–2942, 1981.

[181] G. Rutsch, R. P. Devaty, W. J. Choyke, D. W. Langer, and L. B. Eowland. Measurement of the Hall scattering factor in 4H and 6H SiC epilayers from 40 to 290 K and in magnetic fields up to 9 T. *J. Appl. Phys.*, 84:2062–2064, 2004.

[182] C.-S. Chang, H. R. Fetterman, and C. R. Viswanathan. The characterization of high electron mobility transistors using Shubnikov-de Haas oscillations and geometrical magnetoresistance measurements. *J. Appl. Phys.*, 66:928–936, 1989.

[183] L. Donetti, F. Gámiz, and S. Cristoloveanu. A theoretical interpretation of magnetoresistance mobility in silicon inversion layers. *J. Appl. Phys.*, 102:013708–1 – 013708–6, 2007.