2006

Schottky Field Effect Transistors and Schottky CMOS Circuitry

Reinaldo Vega

Schottky Field Effect Transistors and Schottky CMOS Circuitry

By

Reinaldo A. Vega

A Thesis Submitted

in Partial Fulfillment

of the Requirements for the Degree of

Master of Science

in Microelectronic Engineering

Approved by:

Professor _____
Dr. Karl D. Hirschman (Thesis Advisor)

Professor _____
Dr. Santosh K. Kurinec (Thesis Committee Member)

Professor _____
Dr. Sean L. Rommel (Thesis Committee Member)

Professor _____
Dr. James E. Moon (Thesis Committee Member)

DEPARTMENT OF MICROELECTRONIC ENGINEERING

COLLEGE OF ENGINEERING

ROCHESTER INSTITUTE OF TECHNOLOGY

ROCHESTER, NEW YORK

APRIL 2006

Schottky Field Effect Transistors and Schottky CMOS Circuitry

By

Reinaldo A. Vega

_____ _____

Name Date

Acknowledgment

To begin, I would like to thank my primary advisor, Dr. Karl Hirschman, as well as those on my advisory committee – Dr. Santosh Kurinec, Dr. Sean Rommel, and Dr. James Moon. I am infinitely grateful for their support, for their encouragement, and for all of the stimulating discussions we have had during my time at RIT. I also wish to thank them for advocating my candidacy for various Ph.D. programs and research fellowships, without which my "next step" may very well have been quite different. It is not often that a group of people allows someone the freedom to run with a vision as far as I have tried to do with mine. To call this lucky on my part is an understatement. I would also like to thank Dr. Varda Main of RIT's Technology Licensing Office for all of her help regarding the intellectual property aspects of this work and other manifestations of my imagination.

I also wish to thank Dr. Lynn Fuller and Phu Do for their invaluable contributions to processing techniques and recipes, without which this project would most assuredly have taken much longer. Furthermore, I wish to thank Stephen Sudirgo, Robert Manley, Michael Aquilino, Michael Latham, Vee Chee Hwang, Eric Woodard, Robert Mulfinger,, Daniel Jaeger, and many other colleagues, regarding processing, device physics, simulations, and just about everything else.

Additional thanks go to Charles Gruener, who has been so helpful with mask writing and system administration, as well as Paul Mezzanini and Richard Tolleson for their assistance with the VLSI lab. Crucially important are the contributions from Scott Blondell, David Yackoff, John Nash, Sean O'Brien, Bruce Tolleson, Richard Battaglia, and Thomas Grimsley for their support throughout all of the processing challenges I have encountered in the Semiconductor and Microsystems Fabrication Laboratory (SMFL). I would also like to thank the Microelectronic Engineering Department and its affiliates for all donations relevant to enabling this project.

This page would be far from complete if I did not thank Dr. Michael Potter. Throughout my time at RIT, he has been something of a mentor, without whom I may not have developed the motivation to pursue a Ph.D. There are few if any others who have been more encouraging or more insistent on their opinion of my potential, and his confidence, knowledge, and forward-looking attitude has been most inspiring. Money cannot buy the positive effect that he has had on my growth as an individual.

Most importantly, I am dedicating this work to the visionaries and idealists, those with the passion and the bravery to carve their own independent path in an effort to effect positive change, of this generation and of generations to come. Antiquated modes of thinking do not solve the problems that they create. We must therefore look boldly to the future with awe and wonder as we appreciate what came before us and marvel at the potential of the human mind. It is my hope that the reader will find the presented work a journey in the pursuit of knowledge in the same sense that I have experienced it for myself.

Abstract

It was the primary goal (and result) of the presented work to empirically demonstrate CMOS operation (i.e., inverter transfer characteristics) using metallic/Schottky source/drain MOSFETs (SFETs – Schottky Field Effect Transistors) fabricated on silicon-on-insulator (SOI) substrates – a first-ever in the history of SFET research. Due to its candidacy for present and future CMOS technology, many different research groups have explored different SFET architectures in an effort to maximize performance. In the presented work, an architecture known as a "bulk switching" SFET was fabricated using an implant-to-silicide (ITS) technique, which facilitates a high degree of Schottky barrier lowering and therefore an increase in current injection with minimal process complexity. The different switching mechanism realized with this technique also reduces the ambipolar leakage current that has so often plagued SFETs of more conventional design. In addition, these devices have been utilized in a patent pending approach that may facilitate an increase in circuit density for devices of a given size. In other words, for example, it may be possible to achieve circuit density equivalent to 65 nm technology using a 90 nm process, while at the same time preserving or reducing local interconnect density for enhanced overall system speed. Fabrication details and electrical results will be discussed, as well as some initial modeling efforts toward gaining insight into the details of current injection at the metal-semiconductor (M-S) interface. The challenges faced using the ITS approach at aggressive scales will be discussed, as will the potential advantages and disadvantages of other approaches to SFET technology.

Table of Contents

List of Tables

List of Figures

List of Acronyms

BBT……band-to-band tunneling
BOX……buried oxide
C-V…….capacitance-voltage
CMOS……complementary MOS
CNTFET……carbon nanotube FET
DIBL……..drain-induced barrier lowering
DIST……...drain-induced source tunneling
EOT……equivalent oxide thickness
F-D……..Fermi-Dirac
F-N……..Fowler-Nordheim
FD SOI……...fully depleted SOI
FID……..field-induced drain
FET…….field effect transistor
FUSI……...fully silicided
GIDL……..gate-induced drain leakage
I-V……...current-voltage
IGFET……insulated gate field effect transistor
ILD…….inter-layer dielectric
ITRS……...International Technology Roadmap for Semiconductors
ITS……..implant to/through silicide
LDD……....lightly doped drain
LOCOS……..local oxidation of silicon
LPCVD……..low pressure chemical vapor deposition
M-B……Maxwell-Boltzmann
M-S……metal-semiconductor
MIGS…….metal-induced gap states
MOS……...metal oxide semiconductor
MOSFET……...metal oxide semiconductor field effect transistor
MSD……...metallic source/drain
MSD MOSFET…….metallic source/drain MOSFET
NDR……...negative differential resistance
NFET……..n-channel FET
NMH……..noise margin high
NML……...noise margin low
PD SOI……...partially depleted SOI
PFET……..p-channel FET
POI…….polysilicon-on-insulator
RIT…….Rochester Institute of Technology
RTA……rapid thermal anneal
SB……...Schottky barrier
SB MOSFET…….Schottky barrier MOSFET
SBH……Schottky barrier height
SBL……Schottky barrier lowering
SBTT……..Schottky barrier tunnel transistor

SCE……short channel effect
SDE……source/drain extension
SEM……...scanning electron microscope
SFET…….Schottky field effect transistor
SIIS……silicidation-induced impurity segregation
SIMS…….secondary ion mass spectroscopy
SMFL……Semiconductor and Microsystems Fabrication Laboratory
SOI…….silicon-on-insulator
SRAM……static random access memory
SRIM…….Stopping and Range of Ions in Matter
SRP……spreading resistance profiling
SS……...subthreshold swing
SSD MOSFET……..Schottky source/drain MOSFET
STI…….shallow trench isolation
T-M……...Terada-Muta
TEOS……Tetraethyl Orthosilicate
TFT……thin film transistor
TRIM……Transport of Ions in Matter
UTBSOI……ultrathin body SOI
VTC……...voltage transfer characteristic
WKB…….Wentzel-Kramers-Brillouin

Chapter 1

Introduction: Moore's Self-Fulfilling Prophecy

1.1. Introduction

In 1965, Intel co-founder Gordon Moore made the observation-based prediction that the transistor count on a microchip will double approximately every two years. Over the course of time, his prediction proved accurate, and was since known as "Moore's Law." One might argue, though, that the observation that led to this prediction was from a time when microchip technology was very young, and so the increase in microchip performance was largely for the purpose of advancing the technology in said arena. The formal adoption of Moore's Law by the semiconductor industry as a marketing tool, however, transformed this "prediction" of the rate of technology advancement into an obligation to the customer, who relied (and continues to rely) on the trend of Moore's Law for his or her own business and business models. It would not be unreasonable, then, to suggest that Moore's Law, in modern times, would most appropriately be referred to as Moore's Obligation, or even more fitting, Moore's Self-Fulfilling Prophecy.

Regardless of the mechanisms or "reasons" behind the exponential increase in computing power as a function of time, it nevertheless remains that this trend exists, almost entirely as its own entity with little if any need for encouragement. This trend has driven and continues to drive scientists and engineers from many disciplines to elevate microchip technology to unprecedented levels, only to end up doing it all over again the next year. As a result of this, the microelectronics community continually expresses

concern about the threat of painting themselves into a corner, most usually with regard to image processing capability and the physical limitations of semiconductor devices (and interconnects) on what is approaching the infinitesimal scale.

As a result, while it is the responsibility of the image processing engineer to effectively take ever smaller pictures, the device engineer carries the responsibility of developing an ever superior switch. It is the latter that is the focus of the thesis presented to the reader, but it should be noted that neither of these two challenges is trivial by any stretch of the imagination. It is the intent of the presented thesis, however, to show that complex challenges can be met with elegant solutions.

1.2. Limitations of Conventional CMOS

It should be noted that this section is not intended to discuss conventional CMOS technology in depth, but rather to provide fundamental insight into some of the more "visible" or prevalent challenges of CMOS scaling. That said, on its most simplistic level, switching requires but one thing – energy barrier modulation.

Conventional MOSFETs (or more appropriately, IGFETs – Insulated Gate Field Effect Transistors) achieve switching by modulating what is called a thermal barrier. A simplistic example of this is a "brick wall" for electrons or holes, and modulating the thermal barrier modulates the height of this brick wall. The lower the height, the easier it is for carriers of a given energy to jump over this figurative wall. As this is a *thermal* barrier, if one increases the temperature of the system, more carriers end up existing at a higher energy, and so for a given barrier height, more current flows.

The source/drain regions and the body region of a conventional MOSFET are oppositely doped (i.e., n-type and p-type, respectively, or vice versa), and as a result, a built-in potential barrier occurs at the source-body and drain-body junctions. This built-in potential is the off state thermal barrier to inversion carriers. Ideally, with a grounded gate, modulating the drain-to-source bias (V_{DS}) results in a very low level of current that experiences little if any change. Modulating the gate-to-source bias (V_{GS}), however, modulates the built-in potential, thus modulating the thermal barrier height, thus modulating current flow for a given V_{DS}. This "ideal" behavior is called long channel behavior, and the primary challenge in device scaling is maintaining long channel behavior at very aggressive scales.

For a given device structure (i.e., gate oxide thickness, body doping level, source/drain doping level, source/drain junction depth, etc.), there exists some depletion region width that extends from the source/drain regions into the body region. As the channel length is decreased for the same exact structure, said depletion regions constitute an increasing portion of the body region, and so the quasineutral region available for gate modulation is decreased. This effect is known as "charge sharing" [1], and is illustrated in Fig. 1.1, where x_{dmax} is the maximum gate-induced depletion width, x_j is the source/drain junction depth, x_{dj} is the source/drain depletion region, L is the channel length at the surface, and L_1 is the channel length at x_{dmax}. As a result of charge sharing, gate control over the thermal barrier (and thus current flow through the transistor) becomes gradually replaced by drain control. This is known as the short channel effect (SCE), which manifests itself in the form of drain-induced barrier lowering (DIBL), and

in the most extreme case, punchthrough. DIBL is illustrated in Fig. 1.2, where L_{short} and L_{long} are short and long channel lengths, respectively.

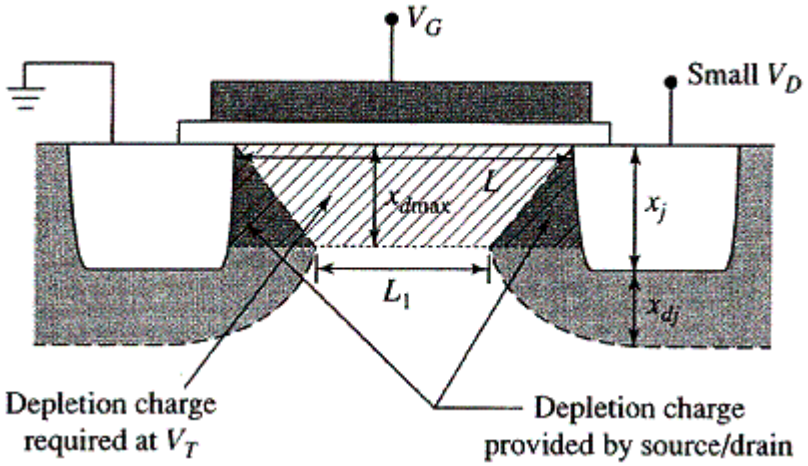

Fig. 1.1. Illustration of charge sharing in a conventional MOSFET, adapted from [1].

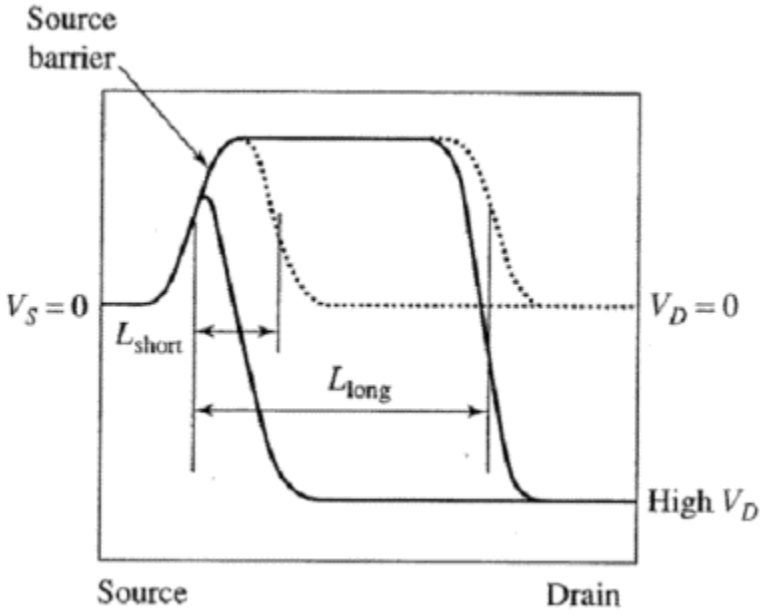

Fig. 1.2. Conduction band of a long and short channel MOSFET, adapted from [1]. For short channel MOSFETs, DIBL reduces the off state thermal barrier at the source.

There are a number of methods one can employ to reduce SCE in aggressively scaled MOSFETs, the most notable of which are to decrease the effective gate oxide

thickness (*EOT*), decrease the source/drain junction depth, and increase the dopant concentration in the body region. Each approach has its own advantages and poses its own challenges.

The effect of decreasing *EOT* is that the gate capacitance increases. As such, for an incremental change in applied gate bias, the incremental change in charge at the semiconductor-oxide interface (and thus the incremental change in surface potential) increases. This is most simply understood by the relationship between charge, capacitance, and voltage: Q = C*V. For MOSFETs at sub-100 nm scales, *EOT* must be very small, on the order of 10 Å or lower, and the transistors are operated at about 1 V. For conventional SiO_2, the consequent electric field across the gate dielectric becomes large enough such that direct tunneling leakage (through a rectangular barrier) turns into Fowler-Nordheim (F-N) tunneling (through a triangular barrier) through the dielectric, thus adding a gate leakage current component to the total device leakage current. One possible solution here is to utilize gate dielectrics with a high dielectric constant (a.k.a. high-k), for which thicker films can be used to achieve the same gate capacitance. However, high-k gate dielectrics pose two challenges – interface charge (they are deposited rather than grown) and process temperature compatibility.

Another solution for decreasing *EOT* is to employ metallic gates rather than polysilicon gates. As a bias is applied to a polysilicon gate, depletion occurs at the oxide-polysilicon interface (known as poly depletion). This depleted polysilicon effectively places a capacitor in series with the capacitance offered by the gate dielectric, thus decreasing the total gate capacitance, which increases *EOT*. By utilizing a metallic gate,

poly depletion is eliminated (depletion regions do not exist in metals), and so the only component of *EOT* is that offered by the gate dielectric.

Decreasing the source/drain junction depth effectively reduces charge sharing. While Fig. 1.1 illustrates charge sharing as an overlap of depletion regions, another way to look at it is to envision some distribution of ionized dopant atoms about the perimeter of the source/drain junctions and some charge distribution of the same polarity at the gate, as Fig. 1.3 shows. Each charged particle "controls" one particle of depletion/inversion charge, and so as the source/drain junction depth increases, the charge controlled by the source/drain regions increases. This results in a more trapezoidal-like region in the body where gate control exists, which increases sub-surface leakage.

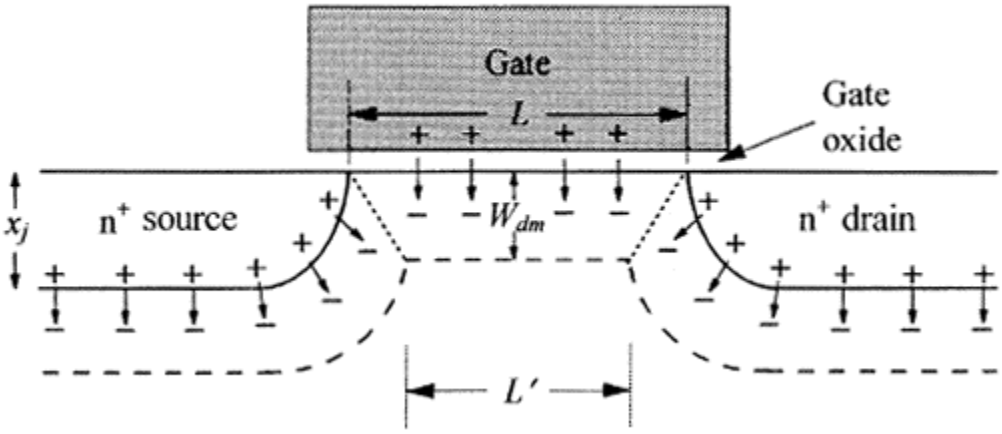

Fig. 1.3. Illustration of charge sharing from the perspective of charge distribution, adapted from [2]. Dotted lines represent the edges of the depletion regions.

In order to decrease the source/drain junction depths, lower energy ion implantation, other implantation techniques (such as plasma implant, pre-amorphization implants, or angled implants to minimize channeling), and rapid thermal annealing (a.k.a.

spike annealing) are usually employed. For aggressively scaled MOSFETs, however, achieving shallow junctions with high junction abruptness and low defect-induced leakage is very challenging.

Increasing the dopant concentration in the body region serves to reduce source/drain depletion region propagation into the body, thus reducing charge sharing and therefore leakage. This comes at a cost, however, as higher dopant concentrations increase coulombic scattering, consequently decreasing channel mobility and hence drive current. Also, higher dopant concentrations in the body region results in a larger threshold voltage, thus decreasing drive current at a given gate and drain bias. This can be engineered around, to some extent, by utilizing a combination of halo/pocket implants and LDD/SDE (lightly doped drain / source/drain extension) regions.

The halo/pocket implants are of the same dopant type as the body region, but of a higher concentration, the goal of which being to confine the source/drain depletion region within the halo/pocket regions. In doing so, the lighter doped portion of the body region remains quasineutral, and thus under greater gate control, with the added effect of increased channel mobility (better drive current); however the drawback here is a larger threshold voltage due to the higher doped halo/pocket regions. This is where the SDE comes in. The use of a SDE ultimately results in a “two stage” source/drain region. The first “stage” is of a relatively low doping (i.e., lower doped than the second stage) with a shallow junction depth, which allows for a somewhat lower halo/pocket dopant concentration. This is the SDE, and it usually interfaces directly with the halo/pocket region. The lower SDE doping decreases the source/drain depletion region propagation into the body region near the surface, thus increasing gate control over inversion charge.

An additional virtue of the SDE region is a reduction in the lateral electric field near the drain, which reduces impact ionization and hot carrier injection into the gate dielectric, as well as gate-induced drain leakage (GIDL – band-to-band tunneling from the drain to the body region induced by gate-to-drain coupling).

The second "stage" of the source/drain region has a high dopant concentration to minimize contact resistance to the metal or metal silicide it comes into contact with, and is recessed behind the SDE region. The degree of recess is controlled by the width of the sidewall spacers surrounding the gate. This recess of the second "stage" increases the effective sub-surface channel length, thus reducing sub-surface leakage. The device structure of this modern conventional MOSFET is illustrated in Fig. 1.4.

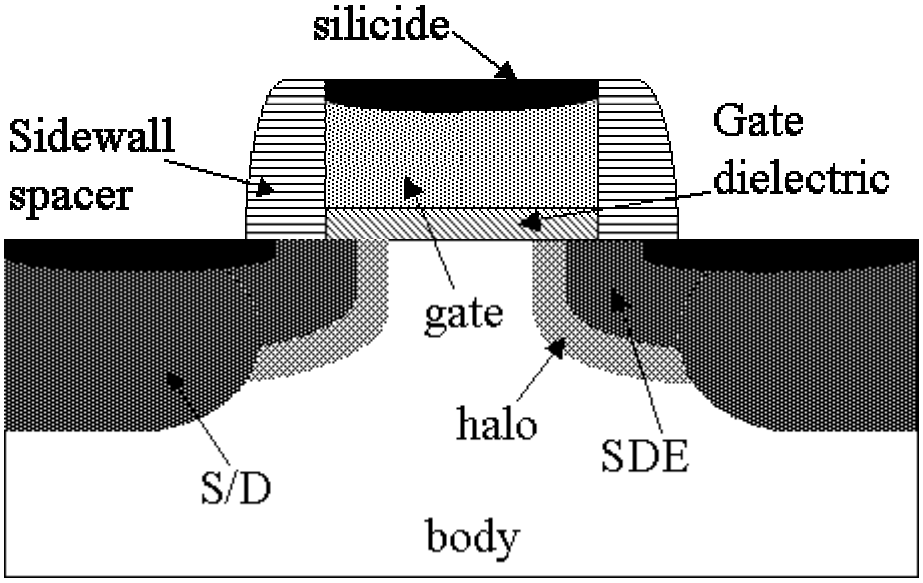

Fig. 1.4. Illustration of a modern MOSFET. The placement of the halo region is non-specific – it can overlap the SDE or lie below it, depending on the preference of the device engineer.

A particular challenge imposed with the introduction of the SDE region is that of series resistance. With a lower dopant concentration than the recessed source/drain region, the series resistance through the SDE region is larger, thus reducing drive current.

In addition, due to the higher source/drain series resistance, V_{GC} (gate-to-channel voltage, which is different from the gate-to-source voltage, V_{GS}) decreases, thus decreasing the level of inversion charge at a given gate bias, thus decreasing drive current. This can be mitigated by reducing the length of the SDE region, which is done by reducing the size of the sidewall spacers. However, this requires caution, as lateral diffusion of the metal silicide used as a source/drain contact can short through the SDE region to the body region, thus increasing leakage.

In considering design criteria of the modern conventional MOSFET, it becomes clear that the engineer is left with many variables to tweak in order to achieve optimal performance. One might argue that the complication involved in engineering sub-100 nm conventional MOSFETs is unnecessary, and that perhaps there exists a more elegant solution. One possible solution is the Schottky Field Effect Transistor, or SFET.

1.3. SFETs as a Solution

Whereas the conventional MOSFET uses source/drain regions made of silicon, the SFET uses source/drain regions made of metal or metal silicide (preferably the latter, as will be shown in subsequent chapters). Instead of forming a conventional thermal barrier at the source-body and drain-body junctions, then, a Schottky barrier is formed, which is a combination of a thermal barrier and a quantum mechanical tunneling barrier. The SFET was first proposed and demonstrated by Lepselter and Sze in 1968 [3] as a p-channel device with PtSi source/drain regions to an n-type body region. The inversion carrier (hole) Schottky barrier height was measured to be ~ 0.25 eV (determined from a current-voltage measurement of a Schottky diode using PtSi to p-type silicon), which

implies an electron barrier height of ~ 0.85 eV. The drive current at low temperatures (77 K) was thought to be due to hole tunneling through the source-body junction, and throughout the history of the SFET, it has been thought that tunneling current is a dominant current mechanism under strong inversion; however, it will be shown in Chapter 4 that the inclusion of Schottky barrier lowering into the device model suggests that thermal current is the dominant current mechanism for low Schottky barrier heights.

Regardless, the first order advantage of the SFET is the simpler fabrication required to achieve the device structure. For SFETs of the simplest, most conventional design, the only implant performed is that which determines the dopant concentration of the body region – recall that the source/drain regions are metallic. Potentially, there exists no need for halo/pocket and SDE implants. Thus, thermal processing is much simplified, and on the surface at least, it is perceived to be much simpler to engineer this device to an optimal level of performance than a conventional MOSFET.

Since the source/drain regions in an SFET are metallic, as opposed to doped silicon, the series resistance of SFET source/drain regions is very low, which should increase drive current and gate control over inversion charge if the Schottky barrier is small enough – the Schottky barrier results in a contact resistance to the body region which effectively replaces the SDE series resistance. In addition to this potential advantage, the metallic source/drain regions form a highly abrupt junction to the body region. This reduces depletion of the net dopant concentration in the body region near the source/drain regions (which is normally a gradient in conventional MOSFETs), which reduces source/drain depletion region propagation through the body, which improves SCE immunity. The basic structure of an SFET is shown in Fig. 1.5.

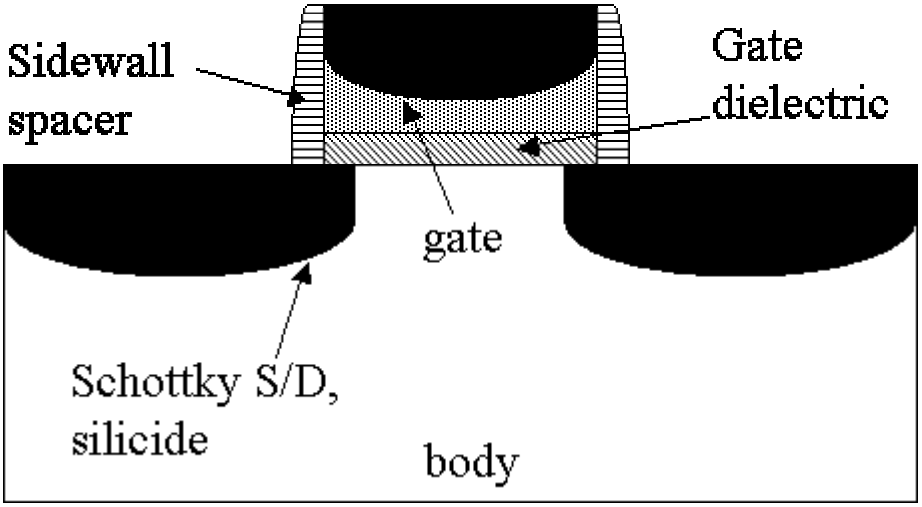

Fig. 1.5. Illustration of a basic SFET structure, with silicide source/drain regions.

Chapter 2 discusses the basic theory behind metal-semiconductor junctions. In Chapter 3, this metal-semiconductor junction theory is expanded in the application to SFETs, whose theory of operation is explored in more detail. Preliminary work at mathematical modeling for SFETs is discussed in Chapter 4, where the focus of the model is to gain insight into the device behavior (with a focus on current transport mechanisms) and to try and quantify design spaces for various design parameters. In Chapter 5, additional attention is given to studying a form of SFET called a bulk-switching SFET. Numerical modeling using SRIM, TRIM, and Silvaco Athena are used to study the structure of the bulk switching SFET and some of the design tricks that can be employed with such a device. Chapter 6 discusses the main points of the device and circuit design implemented for this thesis, as well as the principal advantages that this particular implementation of SFET technology provides for integrated circuit designers and manufacturers (e.g., Fig. 1.6).

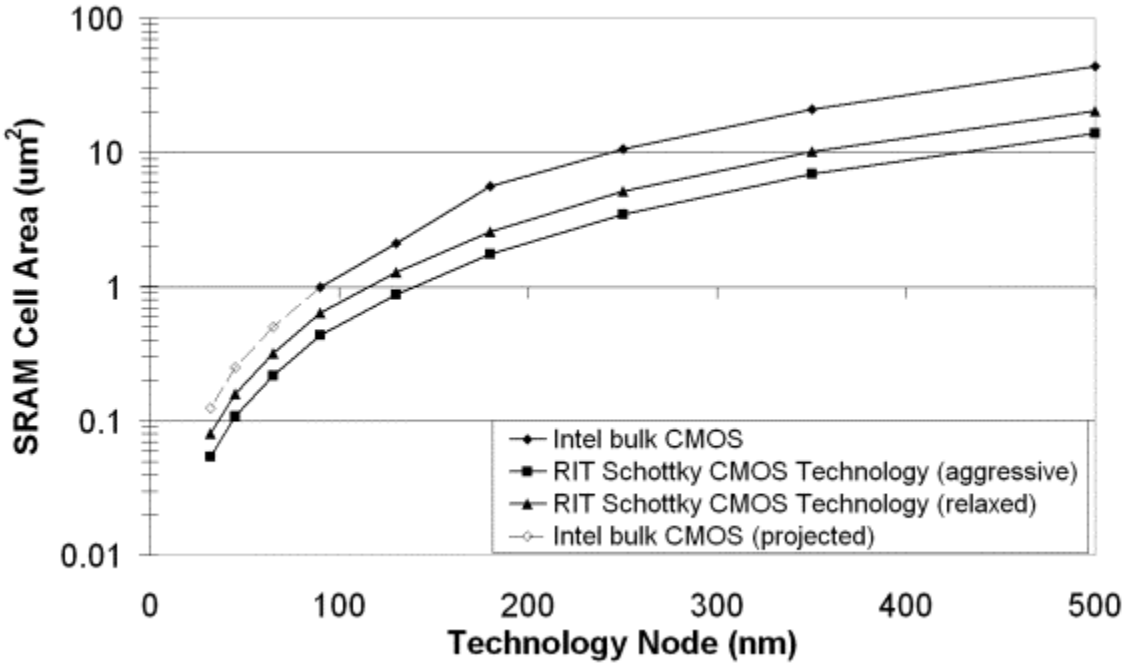

Fig. 1.6. Projected circuit density potential of single metal Schottky CMOS technology, compared to Intel's conventional bulk CMOS with respect to SRAM cell area.

Chapter 7 explores the experimental results of bulk switching SFETs on silicon-on-insulator (SOI) substrates, which includes, to the best of the author's knowledge, the first-ever CMOS demonstration with these devices on SOI (Fig. 1.7). Chapter 8 discusses some anomalous behavior that has been observed in SFETs of conventional design, namely the emergence of a negative differential resistance (NDR) in the I_{DS} vs. V_{DS} characteristic. Chapter 9 explores some initial work performed on bulk-switching SFETs on polysilicon-on-insulator (POI) substrates with, again, an empirical demonstration of CMOS. Chapter 10 concludes this study with a wrap-up of the work performed, the implications of this study for the future potential of SFET technology, and a brief description of some of the more important remaining questions that must be

diligently investigated before this technology can be shown to be a truly viable alternative to conventional CMOS.

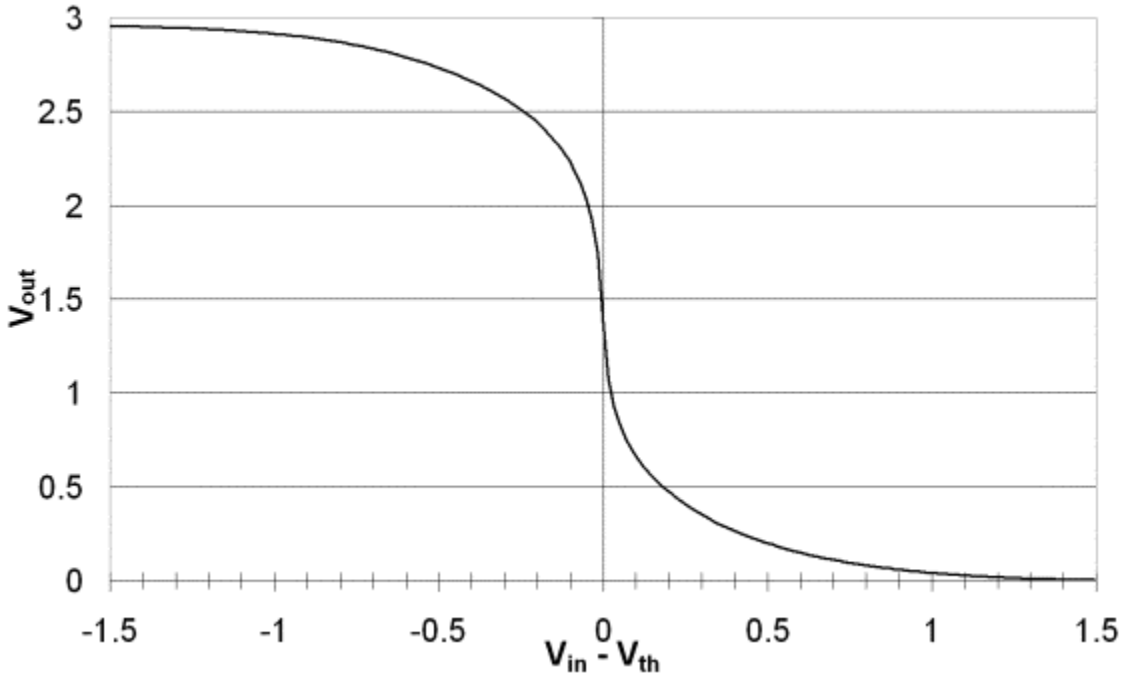

Fig. 1.7. Demonstration of single metal Schottky CMOS on SOI substrates. V_{DD} = 3 V.

Chapter 1 References

[1] R.S. Muller, T.I. Kamins, M. Chan, "Device Electronics for Integrated Circuits, Third Edition," *John Wiley & Sons, Inc.*, 2003, pp. 448, 452.

[2] Y. Taur, T.H. Ning, "Fundamentals of Modern VLSI Devices," *Cambridge University Press*, 1998, p. 142.

[3] M.P. Lepselter, S.M. Sze, "SB-IGFET: An Insulated-Gate Field Effect Transistor Using Schottky Barrier Contacts for Source and Drain," *IEEE Proceedings Letters*, 1968, pp. 1400-1402.

Chapter 2

Metal-Semiconductor Junctions and Schottky Diodes

2.1. Metal-Semiconductor (M-S) Junctions

Unlike a semiconductor, which has a valence energy band edge (E_v), conduction band edge (E_c), and a Fermi level (E_F) between E_c and E_v, metals are treated as only having a Fermi level. The "workfunction" of a given metal is expressed as the difference between E_F and the vacuum level (E_0), and varies from metal to metal (likewise for metal silicides, metal germanides, etc.). Different metals exhibit different workfunctions, and so when placed in direct contact with a semiconductor (which exhibits its own workfunction that is dependent upon the doping level), a built-in potential results (similar to that of a p-n junction, albeit for a different reason). This built-in potential is expressed as the difference between the metal and semiconductor workfunctions, and is ideally greater than zero for electrons when the metal workfunction is larger than the semiconductor workfunction, and for holes when the metal workfunction is smaller than the semiconductor workfunction. While a Schottky diode to electrons or holes is, ideally, only formed under the first or second of the aforementioned conditions, respectively [1], it is important to note that, in all cases, a Schottky *barrier* to *both* carriers is formed in *both* cases.

The barrier height of a Schottky diode is determined by the energy difference between the conduction [valence] band and the Fermi level at the M-S interface for a n-type [p-type] semiconductor. For n-type silicon, for example, if the metal workfunction is larger than the silicon workfunction (i.e., a Schottky diode to electrons), and if the

difference between the metal workfunction and the silicon electron affinity is 0.5 eV, then a Schottky diode is formed with an electron barrier height of 0.5 eV. If for n-type silicon the metal workfunction is lower than the silicon workfunction, then the M-S junction is an ohmic junction for electrons under a forward bias condition (as a Schottky barrier still exists, though, electrons do not experience an ohmic contact under reverse bias in this example). An example Schottky diode to electrons is illustrated in Fig. 2.1, and an example Schottky ohmic contact to electrons is illustrated in Fig. 2.2. In both figures, Φ_M is the metal workfunction, ϕ_B is the Schottky barrier height, ϕi is the built-in voltage of the Schottky diode, Φ_S is the semiconductor workfunction, X is the electron affinity of the semiconductor, ϕ_n is the difference between the conduction band edge and the Fermi level, x_d is the extent of the non-bulk region (where the electric field exists), and L_D is the Debye length (a characteristic measure of the extent of electric field penetration into the semiconductor).

.

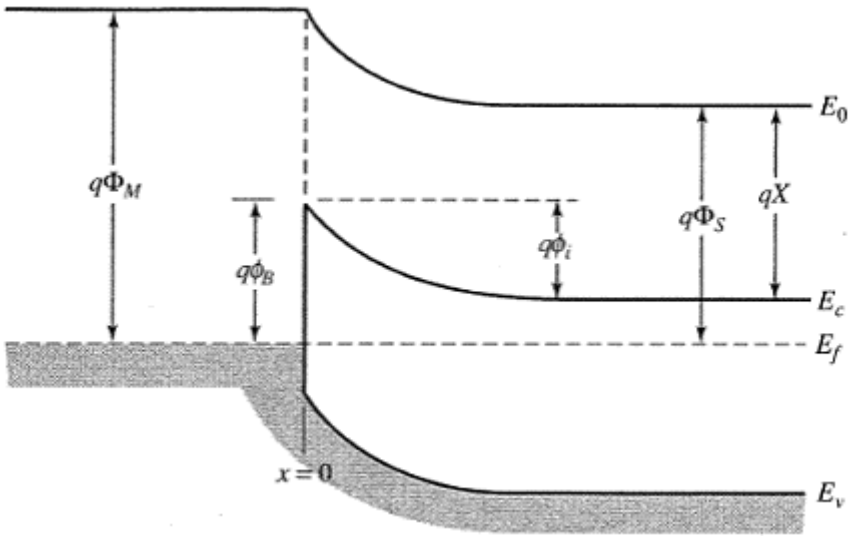

Fig. 2.1. Illustration of a Schottky diode to electrons, adapted from [2].

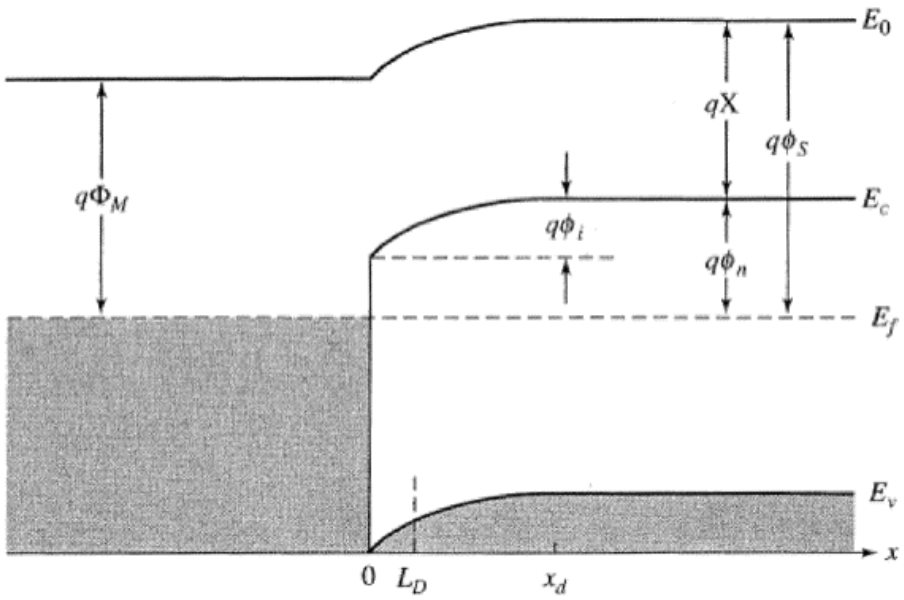

Fig. 2.2. Illustration of a Schottky ohmic contact to electrons, adapted from [2].

Schottky diodes are considered to be majority carrier devices. When the diode is forward biased, majority carriers are injected from the semiconductor to the metal, while minority carriers are injected from the metal to the semiconductor, and the opposite occurs under reverse bias. In either biasing condition, the majority carrier current is always significantly greater than the minority carrier current. In the case of Fig. 2.1, under forward bias, the energy barrier presented to electrons is merely the built-in voltage of the diode (ϕ_i) minus V_a, the applied forward bias. These electrons can "jump over" the barrier (thermionic or thermal current) or quantum mechanically tunnel through it (field emission or tunneling current). Forward bias current is classically expressed as [1]:

$$I = I_s(e^{\frac{qV_a}{kT}} - 1) \tag{1}$$

where q is the charge of an electron, k is Boltzmann's constant, T is the temperature, and

I_s is the reverse bias thermal current, expressed as [1]:

$$I_s = AA^* T^2 e^{\frac{-\phi_B}{kT}} \tag{2}$$

where A is the area, A^* is the effective Richardson's constant (112 A/cm^2·K^2 for electrons in silicon, 32 A/cm^2·K^2 for holes), and ϕ_B is the barrier height.

Continuing with the example in Fig. 2.1, while the electron barrier is the built-in voltage minus V_a under forward bias, the effective hole barrier is the sum of the built-in voltage and the hole Schottky barrier height, ϕ_{Bp} (i.e., the semiconductor bandgap minus the electron Schottky barrier height, ϕ_{Bn}). As this total barrier height is typically larger than the electron barrier, hole thermal current is comparatively small, and since the barrier width is very large, essentially zero hole tunneling takes place.

If the diode in Fig. 2.1 is reverse biased, then the energy barrier presented to electrons is ϕ_{Bn}. Again, electrons must either jump over the barrier or tunnel through it (tunneling increases as the doping in the semiconductor increases, as the barrier width is consequently smaller due to the increase in band bending). As holes are now injected from the silicon to the metal, they essentially "see" an ohmic junction, as no energy barrier exists to impede current flow; however, as holes are being injected from an n-type semiconductor (in which the hole concentration is necessarily very small), the hole injection is very small, and again electron thermal and tunneling currents dominate. Under both forward and reverse biases, therefore, majority carrier current is the dominant current mechanism in a conventional Schottky diode. It will be shown in subsequent sections regarding SFET inversion carrier leakage currents, however, that such is not always the case.

2.2. Interface Characteristics

As it turns out, simply depositing a metal onto a semiconductor, in all likelihood, will not form an ideal Schottky diode. Surface states at the semiconductor surface of the M-S junction alter the surface potential, thus altering the extent of band bending as well as the dimensions of the Schottky barrier. Reference [3] is perhaps the most inclusive theory of Schottky barrier formation, which discusses metal-induced gap states (MIGS) and their effect on the surface potential at the M-S interface. Before discussing MIGS in detail, though, on a superficial level, the Schottky barrier height depends primarily on the surface potential, and the built-in voltage of a given Schottky diode depends on how this potential differs from the potential in the bulk of the semiconductor (i.e., beyond the depletion region).

For example, consider an n-type block of silicon doped at a level of 10^{15} cm^{-3}. This results in a bulk Fermi energy of about 0.3 eV above the intrinsic Fermi level at room temperature, or a potential (relative to the valence band) of about 0.86 eV. Now suppose that the difference between the Fermi level and the valence band edge, due to some mechanism or combination of mechanisms, ends up being 1 eV in equilibrium. This means that the Fermi level is closer to the conduction band at the M-S interface than it is in the bulk, which results in the energy bands bending *downward* from the substrate to the metal. The most immediate effect is a lowering of the electron Schottky barrier height (and a consequent raising of the hole barrier height). This interface potential changes the "crossover" point between electron and hole rectifying contacts. In this particular example, a rectifying contact to holes results for metals having a workfunction of 0.14 eV greater than what would be required in the ideal case for a given dopant

concentration.

To some extent, interfacial oxides or other interfacial layers at non-silicided M-S junctions (i.e., pure metal, no silicide) are treated as a source of surface states that can alter the surface potential [2], as Fig. 2.3 shows for an n-type semiconductor (the states below the Fermi level are filled with electrons, while those above are empty). This interfacial layer is on the order of a few atomic layers, and so the tunneling resistance is quite low. The density of these surface states, D_s, and their distribution about the semiconductor bandgap end up determining what energy the Fermi band at the interface gets "pinned" to (i.e., the Schottky barrier height's dependence on metal workfunction is substantially reduced).

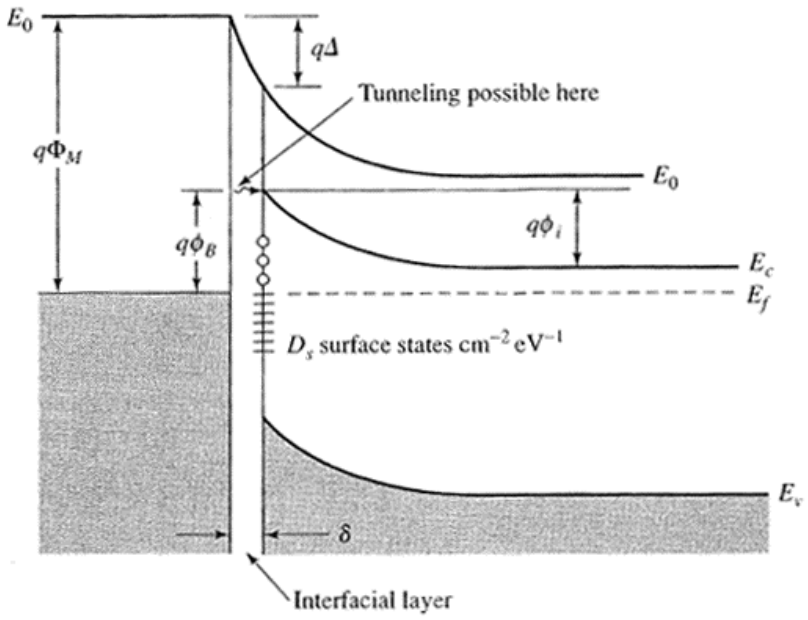

Fig. 2.3. Illustration of M-S junction with an interfacial layer, adapted from [2]. The interfacial layer is treated as the source of surface states, and thus Fermi pinning.

For metal silicides, however, the interfacial layer theory is invalid, as this layer is consumed (and incorporated into the silicide) during the silicidation process, and so the

interface is arguably a “pure” M-S interface. A Schottky barrier still forms, however, which means that interfacial films are not the sole contributor to energy states at the M-S interface (it turns out that interfacial insulators of appropriate quality can actually “de-pin” the Fermi level at the interface [4], but more on this in Chapter 3). Tersoff speculated that the Schottky barrier height has two contributions – a short range contribution, due to surface dipoles, surface bonding details, or M-S electronegativity differences, and metallic screening by MIGS resulting in an additional dipole [3]. It is the metallic screening by MIGS that is considered to result in Fermi pinning. From a wave function perspective, MIGS are Bloch states of the bulk semiconductor with a complex wave vector, or in other words, the decaying tails (known as Heine tails) of the metallic wave functions at varying energies as they propagate into the semiconductor, as Fig. 2.4 illustrates for a single wave function at a single energy level.

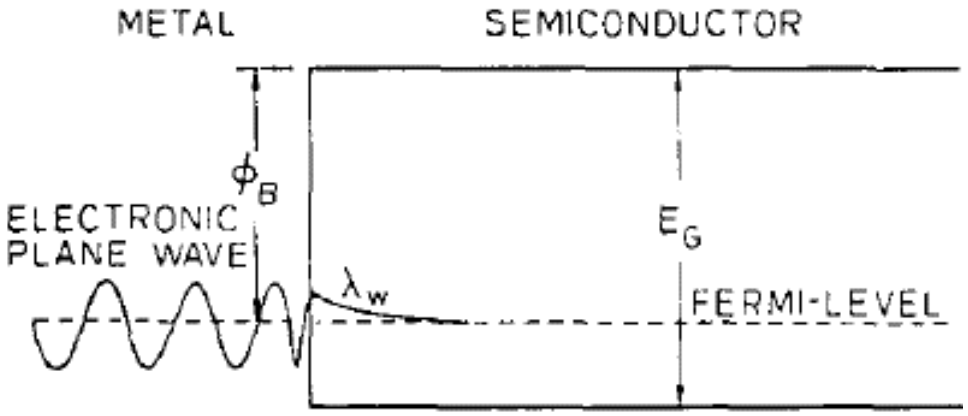

Fig. 2.4. Illustration of Heine tail propagation into a semiconductor, adapted from [5]. λ_w represents the propagation depth, or decay length, of the Heine tail – the probability of an electron existing at some energy within and depth into the barrier.

What results from these Heine tails are a number of gap states close to the interface that “spill over” into the bandgap from the valence and conduction bands, acting

as acceptor-like and donor-like states, respectively. These states have a characteristic distribution within the semiconductor bandgap, which is to say that the bandgap now has a density of states, *N(E)*, at and near the interface. Naturally, these states fill up with electrons and holes, resulting in a carrier distribution throughout the semiconductor bandgap (what was formerly the "forbidden region" of occupation), as Fig. 2.5 illustrates. Within the bandgap, there may exist a minimum value of *N(E)* which corresponds to the energy at which λ_w is the smallest and the gap states cross over from donor-like to acceptor-like behavior. This energy level is called the branch point, E_B, and the Fermi level is pinned precisely at this energy.

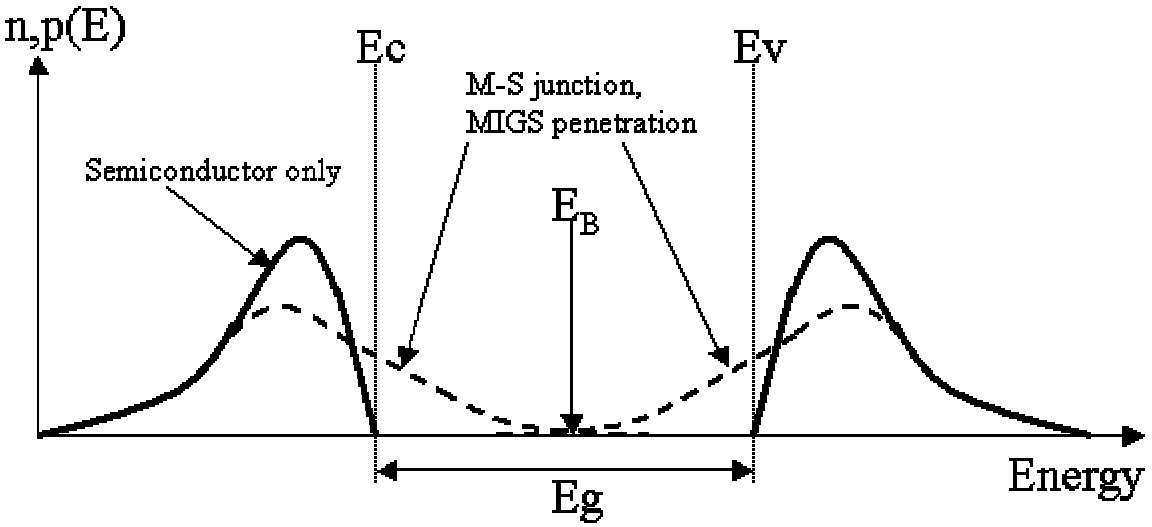

Fig. 2.5. Illustration of carrier concentration vs. energy in the semiconductor only case and a particular M-S junction case, where energy states "spill over" symmetrically from the conduction and valence bands.

Although Fig. 2.5 is a symmetric example that places E_B in the middle of the bandgap, the exact location of E_B may vary between different M-S junctions (which would explain different observed barrier heights with different metals and silicides). For covalent semiconductors such as silicon, germanium, etc., E_B is dependent on surface

state and vacancy levels [3], which accounts for the similar results of different theoretical approaches, such as the interfacial layer approach mentioned previously. As the decay length becomes shorter for larger barrier heights, one might propose that the short range contributions to Fermi pinning mentioned previously end up playing a more significant role for larger M-S electronegativity differences, although the extent towards which this contribution increases is dependent upon the change in decay length with barrier height.

2.3 Schottky Barrier Lowering

At an M-S junction, ionized dopants, defects, and empty or filled gap states provide a charge within the semiconductor. These charges generate electric field lines that terminate at the metal (known as metallic screening) due to an "image" charge that is generated for each charge in the semiconductor that lies within the depletion region or the Heine tail decay length (whichever is larger). This image charge generates an image field whose sign is opposite to the field of the energy bands in the semiconductor near the interface in the ideal case (i.e., a triangular or close-to-triangular energy barrier). The actual potential profile of the Schottky barrier is therefore a function of the superposition of these electric fields, which results in a rounding off of the potential at the top of the barrier. This is known as Schottky barrier lowering or dipole lowering, and is illustrated in Fig. 2.6, where x_m is the position of the peak of the potential profile, $\Delta\phi$ is the change in barrier height due to Schottky barrier lowering, Φ_{B0} is the ideal Schottky barrier height, $E_1(x)$ is the electron energy extending from the metal surface (due to the image force), and $E_2(x)$ is the electron energy extending from the ideal Schottky barrier height into the bulk region, where the electric field is negative.

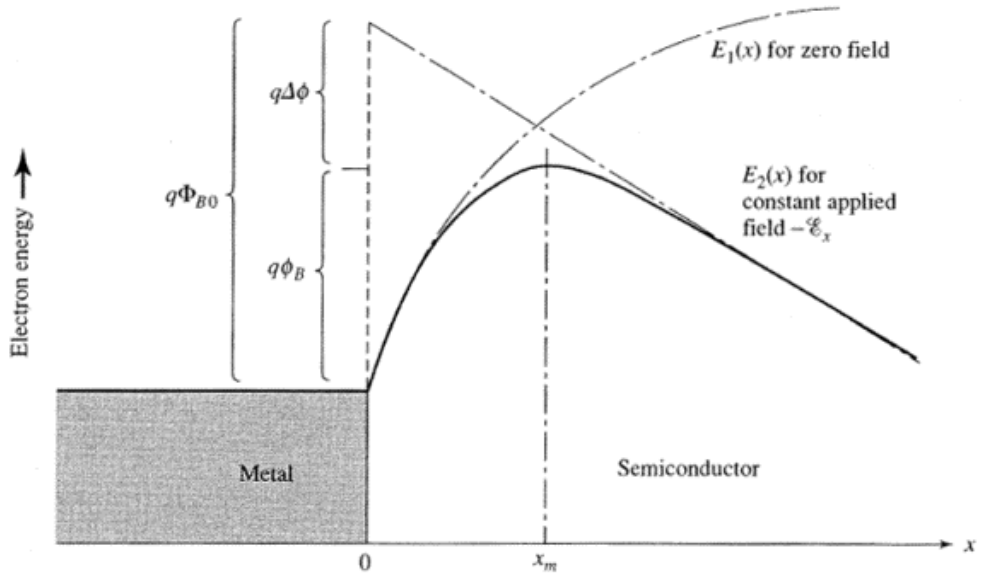

Fig. 2.6. Illustration of Schottky barrier lowering, adapted from [2].

As the electric field at the M-S junction becomes increasingly negative (e.g., increasing reverse bias on a Schottky diode), the extent of barrier lowering increases. Apart from ionized dopants, defects, and gap states, this field can also arise from gate-induced changes in the energy band structure (such as in an SFET or a gated Schottky diode, and it will be shown in Chapter 4 that this barrier lowering effect has implications for SFET design). The expression for Schottky barrier lowering is as follows [1]:

$$\Delta\phi_B = \sqrt{\frac{q|\xi|}{4\pi\varepsilon_0\varepsilon_{si}}} \tag{3}$$

where ε_0 and ε_{si} are the vacuum permittivity and relative permittivity of silicon, respectively, q is the electron charge, and ξ is the electric field at the junction (which must be negative to induce barrier lowering for electrons and positive for holes).

2.4 Quantum Mechanical Tunneling

One of the more interesting properties of the Schottky barrier is that it not only presents a thermal barrier to carriers, like a p-n junction, but it also presents a quantum mechanical tunneling barrier to carriers of insufficient energy to surmount the thermal barrier. What is perhaps the simplest method of visualizing tunneling is to view it from the perspective of wave propagation, illustrated in Fig. 2.7. As the electron wave function approaches an energy barrier of a finite height ϕ_B, one can consider it to have a normalized amplitude of 1. Once the wave function reaches the barrier, its amplitude decays exponentially until reaching the end of the barrier ($x = L$ in Fig. 2.7), at which point the wave function is sinusoidal again, but with reduced amplitude. If the amplitude (normalized with respect to the incoming wave function) of the transmitted wave function is, for example, 0.2, then the tunneling probability for that electron through this particular barrier at this particular energy level is 20%, while the reflection probability is 80%.

One might suppose that "tunneling" has something of a magical or philosophical connotation, but a more appropriate characterization of this phenomenon would be damped transmission – the wave function is a *probability* wave and only collapses into a particular state (i.e., transmitted or reflected) under observation, and the probability of collapsing into that state depends on the amplitude of the wave function in that region. As such, one might consider the electron to be on *both* ends of the energy barrier until the wave function collapses into one state or the other under observation. This is the reality, as well as the oddity, of quantum mechanics, and so one must not make the mistake of concluding that a quantum mechanical phenomenon such as tunneling necessitates a "particle physics-like" explanation. In the application to a large body of electrons,

however, and in this particular example, it can be interpreted that 20% of a given body of electrons traveling toward the energy barrier at that particular energy level will "tunnel" through the barrier, which would give *the same observed result* as the sum of the reduced amplitudes of *all* of the electron wave functions as they exit the barrier (where they are at 20% of their initial value). There would be *zero* difference in measurement between these two interpretations.

The degree or sharpness of exponential decay of the wave function is dependent upon ϕ_B (larger values of ϕ_B result in sharper decay, and for $\phi_B = \infty$, the decay is infinite – no tunneling occurs), and the total transmission is dependent upon ϕ_B and the barrier width, W_B (or L in the case of Fig. 2.7). In the application to Schottky barriers, W_B varies with energy, and carriers closer to the top of the Schottky barrier have a larger probability of tunneling through it.

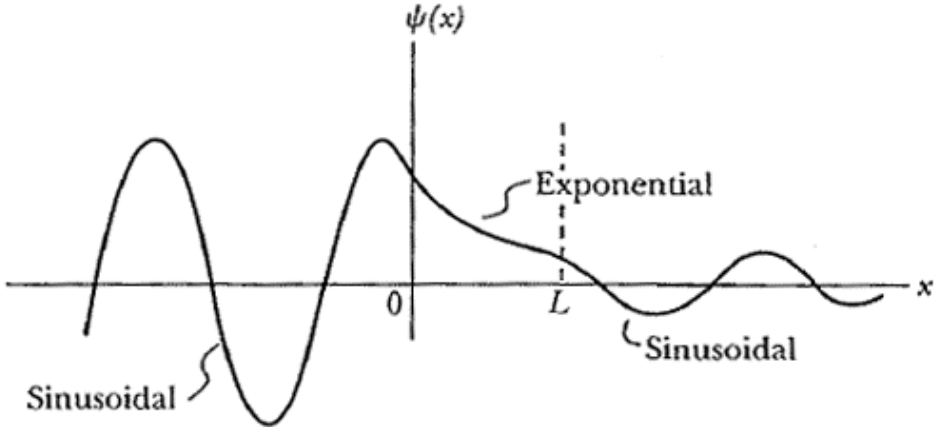

Fig. 2.7. Wave function representation of quantum mechanical tunneling, adapted from [5]. $\Psi(x)$ is the electron wave function, and L is the energy barrier width.

Tunneling current has its most significant contribution to the total current through a Schottky diode when said diode is under reverse bias (when majority carriers are injected from the metal into the semiconductor). Under such conditions, W_B becomes

ever smaller with an increasing electric field, and so the wave function decay takes place over a shorter distance. As such, the normalized amplitude of the transmitted wave function approaches 1 (but never actually reaches 1 due to the physical existence of the barrier, no matter how small W_B is), consequently increasing tunneling current. While it will be shown in Chapter 3 that tunneling current is indeed a component of SFET current, in Chapter 4 it will be shown mathematically that it does not play a dominant role in the on state current of high performance SFETs – thermal current is always larger.

Chapter 2 References

[1] R.F. Pierret, "Semiconductor Device Fundamentals," *Addison-Wesley Publishing Company, Inc.*, 1996, pp. 477-492.

[2] R.S. Muller, T.I. Kamins, M. Chan, "Device Electronics for Integrated Circuits, Third Edition," *John Wiley & Sons, Inc.*, 2003, pp. 144-165.

[3] J. Tersoff, "Schottky Barrier Heights and the Continuum of Gap States," *Phys. Rev. Lett.*, Vol. 52, no. 6, 1984, pp. 465-468.

[4] D. Connelly, C. Faulkner, D.E. Grupp, J.S. Harris, "A New Route to Zero-Barrier Metal Source/Drain MOSFETs," *IEEE Trans. on Nanotechnology*, 2004, Vol. 3, no. 1, pp. 98-104.

[5] S.T. Thornton, A. Rex, "Modern Physics for Scientists and Engineers, Second Edition," *Saunders College Publishing*, 2000, p. 210.

Chapter 3

Schottky Field Effect Transistors – Theory of Operation

3.1. The Schottky Field Effect Transistor (SFET) and Ambipolarity

With metallic source/drain regions in an SFET (also known as an SB MOSFET [Schottky Barrier MOSFET], SBTT [Schottky Barrier Tunnel Transistor], or SSD MOSFET [Schottky Source/Drain MOSFET]), Schottky diodes, as opposed to p-n diodes in conventional MOSFETs, exist which inhibit current flow. By applying a gate bias, band bending of the substrate takes place at the semiconductor-gate dielectric interface, consequently modifying the geometry of the source/drain Schottky barriers. In doing so, the source-to-drain current is modulated by the gate terminal. Unlike conventional MOSFETs, however, SFETs are ambipolar, meaning that two I-V characteristics (NFET-like and PFET-like) are attainable with a single device. Fig. 3.1 loosely illustrates this ambipolarity in the example of an n-body SFET. By applying a positive gate bias, electrons (majority carriers in this case) accumulate at the semiconductor-gate dielectric interface. This increases the degree of band bending at the source/body and drain/body junctions, consequently decreasing the barrier width at each junction. As the barrier width is made smaller, the tunneling probability for the accumulated carriers increases. In accumulation mode, then, and for relatively large Schottky barrier heights to electrons (ϕ_{Bn}), tunneling current is the dominant current component that is gate-modulated. By applying a negative gate bias to this n-body SFET, the body region depletes and eventually inverts to form a p-type channel. If the electron barrier height is large, then the inversion mode (hole) barrier height ($\phi_{Bp} = Eg - \phi_{Bn}$) is relatively small, and so

thermionic emission of holes over the Schottky barrier is the dominant current mechanism. Thus, by simply reversing the polarity of the applied gate bias, the SFET can switch from accumulation mode to inversion mode and back.

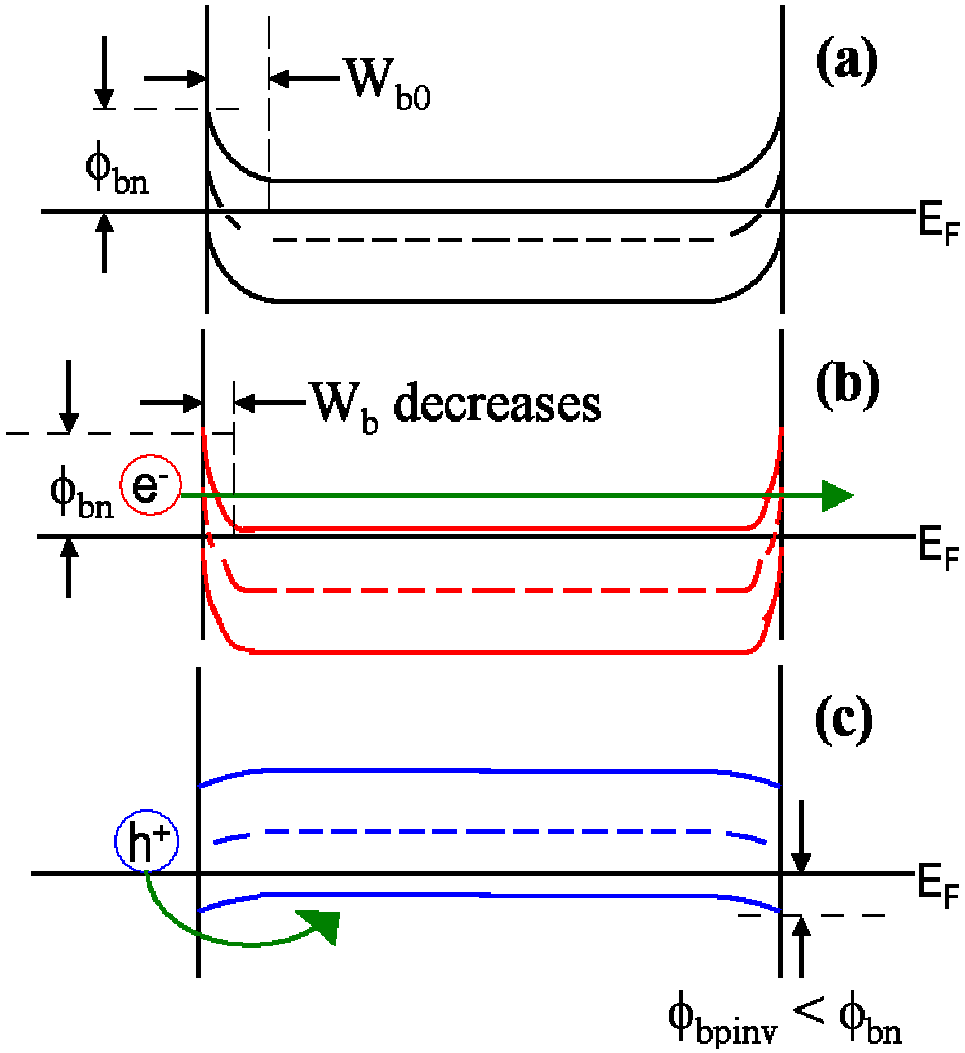

Fig. 3.1. Basic illustration of ambipolar operation in an n-body SFET. (a) n-body SFET in equilibrium; (b) n-body SFET in accumulation mode; (c) n-body SFET in inversion mode.

As it turns out, the SFET is not the only semiconductor device that exhibits such a characteristic. Carbon nanotube transistors, or CNTFETs, have been shown to perform in the same fashion [1], [2]. In fact, not only do CNTFETs exhibit ambipolar behavior, but

the mechanisms of that behavior are exactly the same as for SFETs – Schottky barrier source/drain regions. It then follows that, at least on a superficial level, SFETs and CNTFETs are one and the same, with the only difference being the type of semiconductor used. It is no coincidence, then, that the performance limitation factors of both devices, as well as the optimal design approaches for both devices, are essentially the same [2], [3], [4]. These will be discussed later in further detail.

3.2. Leakage and Drain-Induced Barrier Lowering (DIBL)

With the considerations of gate leakage and defect/trap-induced leakage aside, the leakage mechanisms in an SFET are considerably different from those in a conventional MOSFET, as again the diodes in use are Schottky diodes as opposed to p-n diodes. In a conventional MOSFET, some of the leakage results from generation current in the reverse biased diode (drain-body diode). Another leakage mechanism, particularly in short channel devices, results from drain-induced barrier lowering (DIBL), whereby the electric field from the drain effectively decreases the thermal barrier height at the source/body junction (this can also be viewed as applying a forward bias of some magnitude to the source-body junction). Further leakage can be induced via gate-to-drain coupling at large drain biases and a relatively low or zero gate bias, whereby the magnitude of the surface potential in the body region near the drain is large enough to induce band-to-band tunneling between the drain and the body. Such leakage is called gate-induced drain leakage, or GIDL.

In an SFET, tunneling through and thermal current over the Schottky barriers are the primary leakage mechanisms; however, since SFETs are ambipolar, what is referred

to as “leakage” is typically the device attempting to “turn on” in the other direction. For example, consider the n-body SFET in Fig. 3.1. Inversion mode operation requires a negative gate bias; however, if the gate is grounded and the drain is at some negative bias (V_{GD} is positive), gate-to-drain coupling accumulates electrons at the drain-body junction, thus decreasing the Schottky barrier width. This “leakage” is electron tunneling current through the reverse biased drain-body diode, which is effectively the device turning on or trying to turn on in accumulation mode. Likewise, for the same device operating in accumulation mode, a positive gate bias is required; however, with a grounded gate and a high enough positive drain bias (V_{GD} is negative), the semiconductor near the drain undergoes inversion. Thus, “leakage” for accumulation mode operation has an inversion mode component (hole thermal current in this example). There is also some drain-induced tunneling leakage at the source/body junction (electron tunneling in this example) under the same biasing conditions, whereby the field from the drain, if large enough, will shrink the source/body barrier width, facilitating carrier tunneling injection from the source. One might call this drain-induced source tunneling (DIST). All of these effects can be viewed as DIBL in either a non-conventional sense (i.e., a tunneling barrier width is being modulated as opposed to a thermal barrier height) or in a conventional sense (i.e., thermal barrier modulation), as an increase in drain bias is reducing some sort of energy barrier that would normally impede current flow.

For the biases involved in inversion mode operation and low inversion mode barrier heights, thermal leakage of inversion carriers over the source/body Schottky barrier is of particular importance. To reduce this leakage, the body dopant concentration can be increased. This results in band bending that creates a thermal barrier to inversion

carriers that extends beyond the respective Schottky barrier height. The extent of this band bending is called the contact potential, ϕ_c [5], and is illustrated in Fig. 3.2. The actual barrier height to inversion carriers, then, becomes the sum of the contact potential and the Schottky barrier height.

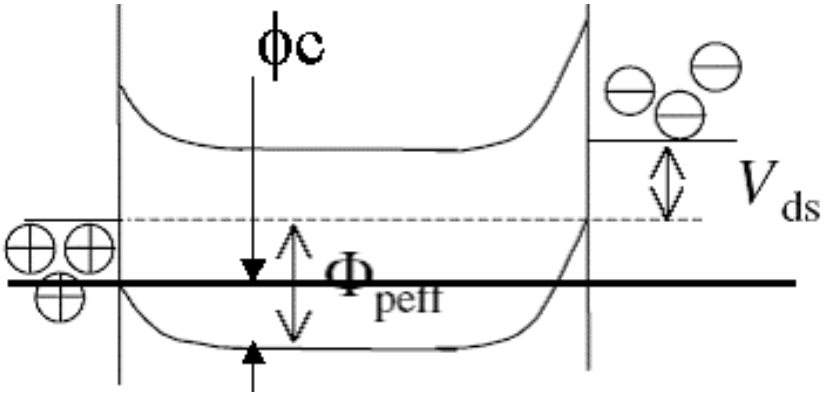

Fig. 3.2. Simplified band structure for an n-body SFET (Schottky "PFET"), adapted from [5] and modified. Φ_{peff} is the total inversion carrier (hole) barrier height.

In increasing the body doping level, however, one must take caution, as there is a consequent decrease in the tunneling barrier width to accumulation carriers at the drain, as Fig. 3.2 also illustrates. While a high body doping (or an appropriate shift in the gate workfunction for a device which primarily exhibits surface channel behavior) can dramatically reduce inversion carrier leakage, accumulation carrier leakage will increase. A mathematical and empirical quantification of this tradeoff will be supplied in subsequent chapters. Another effect of increasing the body doping is that the channel mobility is reduced, which, coupled with the consequent increase in threshold voltage, can decrease drive current. For aggressively scaled devices that exhibit ballistic or near-ballistic transport, however, this may not be an issue. It should also be noted that, unlike the Schottky barriers, ϕ_c will change with channel length (this is a generic assumption,

though, as at very aggressive scales, Fermi pinning at the M-S junction may not be so definite due to the dominance of microstructural effects over macrostructural effects).

3.3. Subthreshold Swing

For an SFET consisting of a body region that is doped to a level so as to result in a contact potential (ϕ_c) to inversion carriers, as the gate bias is modulated in the direction that depletes and eventually inverts the body region, at first only ϕ_c is modulated. The subthreshold behavior in this region of operation is the same as that of a conventional MOSFET, as there is not yet any Schottky barrier modulation. Thus, in this region of operation, the subthreshold swing is purely a function of effective oxide thickness (*EOT*), body dopant concentration (which affects the maximum gate-induced depletion width, W_{Dmax}), and channel length (for short channel devices) [6].

At some gate bias, called the source-body flatband voltage (V_{sbfb}) [7], ϕ_c reaches zero and the valence and conduction bands are flat. Increasing the magnitude of the gate bias beyond this point modulates the Schottky barrier. It has been suggested that only the Schottky barrier width is modulated here, and that any increase in current for V_{GS} beyond V_{sbfb} is due strictly to an increase in tunneling current at the source-body junction [7]-[10]. However, as the electric field at the source-body junction is of the appropriate polarity to induce Schottky barrier lowering ($\Delta\phi_B$), there is reason to suggest that the Schottky barrier height is also modulated for V_{GS} beyond V_{sbfb} [11]. This has important design implications regarding what actually is and is not an acceptable inversion mode Schottky barrier height (and therefore what are and are not acceptable material choices) [11], but the point here is that a different energy barrier is being modulated. As this

different type of energy barrier is being modulated, which exhibits a different sensitivity to changes in V_{GS}, a shift in the subthreshold swing can occur at V_{sbfb} if the inversion mode Schottky barrier height is large enough to effect a change in current flow. Figs. 3.3 and 3.4 illustrate, respectively, conduction band modulation with V_{GS} in a p-body SFET without and with Schottky barrier lowering.

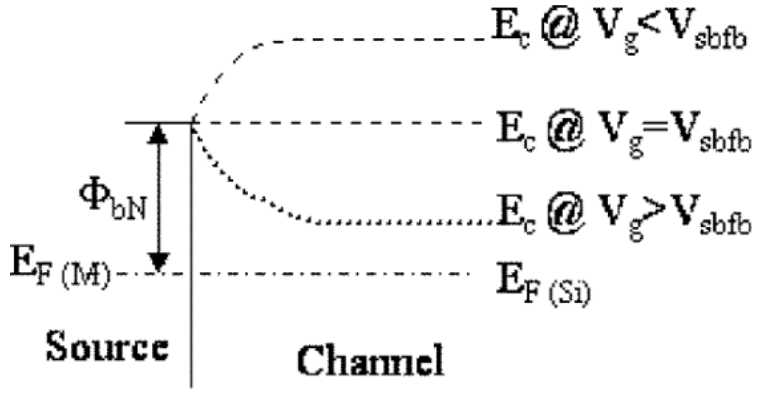

Fig. 3.3. Illustration of conduction band modulation with gate bias in a p-body SFET, according to "conventional" SFET theory, adapted from [7].

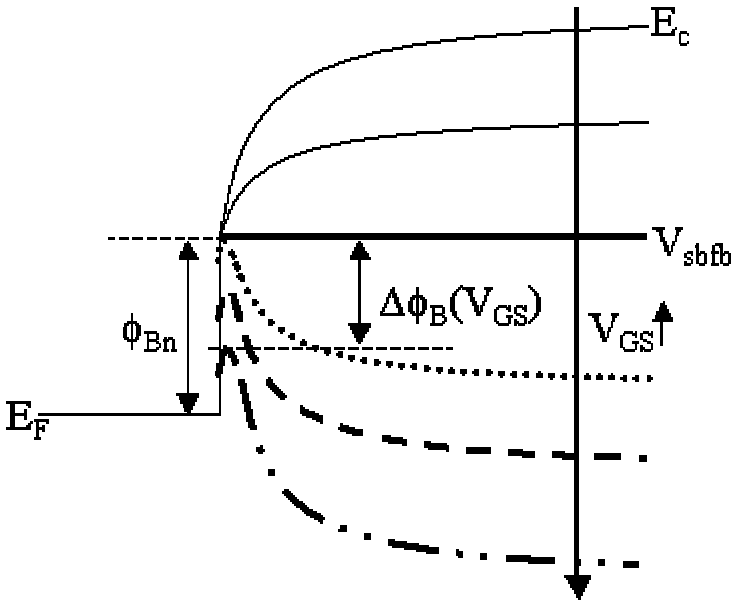

Fig. 3.4. Illustration of conduction band modulation with gate bias in a p-body SFET, accounting for Schottky barrier lowering. In this case, the Schottky barrier becomes narrower *and* shorter for V_{GS} greater than V_{sbfb}.

For a given device structure (i.e., gate oxide thickness, body dopant concentration, etc.), it would follow that the subthreshold slope at V_{GS} beyond V_{sbfb} is independent of Schottky barrier height if thermal current dominates in this region of operation, as $\Delta\phi_B$ is largely independent of ϕ_B. If tunneling current is the dominant fraction of current injection, such as in accumulation mode operation, one might suggest a shift in the subthreshold slope as ϕ_B is modulated; however, $\Delta\phi_B$ also occurs in this mode of operation, and the results in [11] and Chapter 4 suggest that, regardless of whether thermal or tunneling current dominates, the subthreshold slope in the respective region of operation remains relatively constant for different values of ϕ_B as the Schottky barrier is modulated by the gate for a given device structure.

Also, for a given device structure, it is possible to alter V_{sbfb}. To a first order at least, V_{sbfb} is dependent on two factors – the inversion mode Schottky barrier height and short channel effects (i.e., DIBL). A larger barrier height will decrease V_{sbfb}, as would a larger degree of DIBL. Distinguishing between the two in a test environment would simply require testing the same device structure at different channel lengths – DIBL is dependent on channel length, while ϕ_B is not.

3.4. Optimizing SFET Performance

The implications of Schottky barrier modulation versus conventional thermal barrier modulation in an optimized SFET design may, at least in some cases, seem rather obvious – use a small Schottky barrier height and a moderate to high body doping of the appropriate species (or some other means of band bending, such as gate workfunction). In such a case, only ϕ_c modulation exists for most or all of the bias values within the

power supply voltage (V_{DD}). However, quantifying exactly what barrier height is "small enough" and determining what material/s or technique/s would provide for the best approach is not a trivial matter, particularly in the sub-100 nm regime where process variation is greater and silicide linewidth dependence limits material choices.

From a materials perspective, it has been suggested that rare earth metals such as Pt and Er, when used as a silicide, provide for the best to-date inversion mode Schottky barrier heights (~ 0.25 eV) [5], [7], [12]-[15]. Rare earth metals, however, are necessarily rare and expensive, and so while they are useful from a research perspective, they beg the question of long-term sustainability in a high-volume manufacturing environment. It has been shown that silicides formed with less exotic refractory metals (namely NiSi) can be "tuned" to result in very low barrier heights. In particular, S^+ implantation at doses between $1x10^{13}$ and $2x10^{14}$ cm^{-2} before Ni deposition (and subsequent silicidation) was shown in [16] to dramatically alter the electron Schottky barrier height, ϕ_{Bn}, from its zero-point value of 0.65 eV to values as low as 0.07 eV. This is considerably lower than the ~ 0.25 eV provided by $ErSi_2$, and so should prove superior for an n-channel SFET in terms of both performance and practicality. The mechanism behind this effect of barrier height tuning was mentioned in [16] to be due to passivation of the silicon surface with valence-mending adsorbates (i.e., the broken bonds at the silicon surface are "filled," which is best done with a Group VI element such as S or Se). It would seem, then, that the branch point (E_B from Chapter 2) shifts toward the conduction band with higher S^+ doses, due a change in degeneracy within the bandgap, or more appropriately, a shift in the dependence of barrier height toward workfunction differences and away from E_B. Likewise, it would not be unreasonable to suggest a

similar possibility for attaining very low *hole* Schottky barrier heights with NiSi or some other refractory metal silicide (preferably a high workfunction material) to replace PtSi as a candidate for p-channel SFETS (in which case, E_B would have to shift toward the valence band). To date, however, no known work has been performed on this.

Different analyses give different suggestions as to what the maximum barrier height is in order to be competitive with conventional MOSFETs. Some propose barrier heights on the order of 0.1 eV to 0.15 eV [15], while others have gone so far as to propose *negative* barrier heights [8]. Rather than tuning a given material to achieve high performance, it may also be possible to eliminate the Schottky barrier altogether, by "depinning" the Fermi level at the M-S interface. This effectively frees up the surface potential at the M-S interface so that it can be controlled by a gate bias rather than by the metal. This is almost the same exact thing as using valence-mending adsorbates, except that here it is done with an interfacial material.

Such a depinning approach is what was proposed in [15], where a very thin interfacial film is intentionally placed between a metal and a semiconductor (i.e., no silicidation is performed). This film is of sufficient thickness to block MIGS penetration into the semiconductor, but also thin enough to provide minimal tunneling resistance. Fig. 3.5 qualitatively describes this approach. The end result is a M-S junction with minimal contact resistance, and thus optimal drive current. Due to the depinning of the Fermi level, no tunneling barrier is modulated with V_{GS} – only a thermal barrier is modulated, the height of which is the difference between the Fermi level in the metal and the conduction or valence band in the semiconductor (depending on the channel type). In [15], only M-S junctions were fabricated to test and ultimately prove this depinning

concept (Fig. 3.6). Thermal SiN_x was used as the interfacial layer, with an estimated thickness of 1-2 monolayers providing the best contact resistance. As Fig. 3.6 shows, there is more process latitude with interfacial films that are "too thick," as the contact resistance is less sensitive to large growth times versus very small growth times. However, as these junctions use pure metal and not metal silicide, fabricating aggressively scaled SFETs would require a new process or device structure that would facilitate self-aligned gate technology.

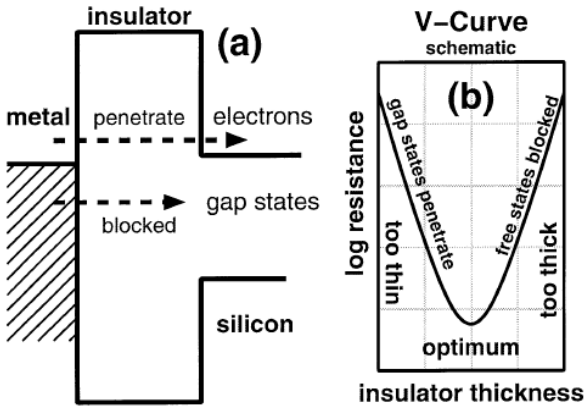

Fig. 3.5. MIGS blocking and the dependence of contact resistance on interfacial layer thickness, adapted from [15]. (a) MIGS penetration blocked by an interfacial layer; (b) a balance between MIGS blocking and tunneling resistance must be reached to achieve optimal performance.

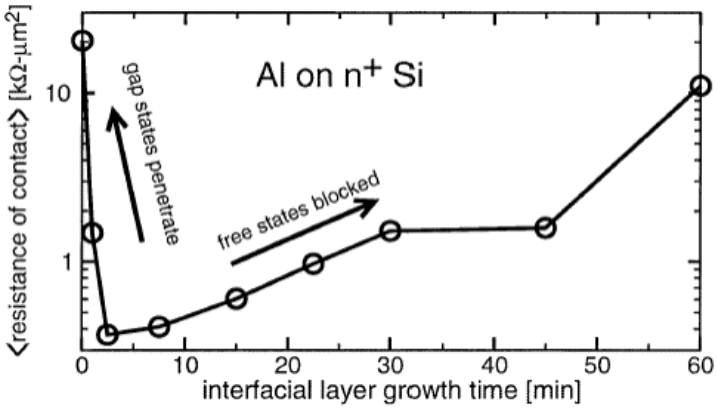

Fig. 3.6. Contact resistance vs. interfacial layer growth time, adapted from [15]. Aluminum was the metal of choice, with heavily-doped n-type silicon as the semiconductor material, and thermal SiN_x as the interfacial layer.

From a techniques perspective, there are a number of methods one can employ to optimize SFET performance. Connelly *et al.* explored the optimization of gate-to-source/drain offset for 25 nm dual-gate metallic and doped source/drain devices, and found that in *both* cases, some degree of gate *underlap* provides the optimal balance between drive current and leakage current [17]. This optimal underlap was found to vary between 0.5 nm and 2 nm for silicon-on-insulator (SOI) devices with body thicknesses between 6 nm and 10 nm, and for inversion mode Schottky barrier heights between 0 eV and 0.2 eV. At such dimensions, source-to-drain coupling becomes significant when gate-to-source/drain overlap exists, which results in an increase in V_T rolloff (SCE). Pulling the source/drain regions farther apart (effectively increasing the channel length, L_{ch}) such that an underlap to the gate exists decreases this coupling while also eliminating overlap capacitance. In addition, V_T rolloff decreases, as does the subthreshold swing. However, this change in subthreshold swing is due to a decrease in conventional SCE. For SFETs, as the inversion mode Schottky barrier height increases, the optimal gate underlap decreases, as the fringing fields from the sides of the gate are not as effective at modulating the Schottky barrier than if there were some degree of gate overlap. Too large a gate underlap, then, while decreasing leakage and reducing SCE, will also reduce drive current due to an increase in series resistance. Not surprisingly, then, one is brought back to the fundamental challenge in designing transistors – the ubiquitous tradeoff between drive current and leakage current.

Knoch and Appenzeller mathematically explored the performance effect on SFETs of using ultrathin body, fully depleted SOI substrates [10]. Their analysis showed that for a small enough body thickness and/or a small enough *EOT*, the electrostatic

effect of the gate causes a significant reduction in the Schottky barrier, thus increasing drive current and improving the subthreshold swing. Although their analysis suggested that tunneling current dominates in the on state, it will be shown in Chapter 4 that the role of tunneling current in the on state has a dependence on the Schottky barrier height.

Although the effect of a small *EOT* is intuitive – better gate coupling to the body region – the effect of thinner body regions with a small *EOT* on current injection is interesting. While thinner body regions can reduce SCE in conventional MOSFETs and SFETs, the channel resistance increases due to a reduced cross-sectional surface area, thus potentially decreasing drive current (though for aggressively scaled devices, ballistic or near-ballistic transport can be assumed). However, unlike conventional MOSFETs, current injection at high drain biases in SFETs is strongly dependent on the entire shape of the energy barrier at the source, and the shape of this barrier has little if any dependence on channel length. This explains why it is possible to observe relatively large subthreshold swings (i.e., poor characteristics) for SFETs with large channel lengths if the Schottky barrier is high enough [7]. So, while decreasing the body thickness may well increase the channel resistance for devices exhibiting non-ballistic transport, it also facilitates an increase in current injection through and over the Schottky barrier at the source. For devices exhibiting ballistic or near-ballistic transport, then, ultrathin SOI (on the order of 4 nm [10]) with fully depleted body regions would seem to optimize drive current, as well as subthreshold swing (and, consequently, transconductance). One must take caution in thinning the body region, however, as eventually quantum confinement within the body will increase the effective Schottky barrier height [7], [8].

Rather than, or perhaps in addition to utilizing ultrathin body regions, valence-mending adsorbates, different silicide materials, and optimizing gate overlap/underlap, one can also engineer the semiconductor bandgap itself to modulate the Schottky barrier height. $Si_{1-x}Ge_x$ is a promising semiconductor, as it is compatible with conventional silicon processes and the bandgap can be reduced by ~ 42 meV per 10% Ge content [18]. While this is not a very large drop in the bandgap, channel mobility (for electrons at least) is dramatically enhanced [19] and the barrier height to a given silicide (or more appropriately, germanide) may change depending on where the MIGS-induced branch point lies. This also applies to pure germanium, which has a bandgap of ~ 0.66 eV [20]. In the case of pure germanium, though, the largest (i.e., worst case) non-modified Schottky barrier for the better channel type (i.e., the one with the lowest inversion mode barrier height) is half the bandgap, or ~ 0.33 eV, which is much smaller than what is available with silicon (~ 0.56 eV), and potentially better for drive current.

However, with a smaller bandgap, and hence a smaller barrier height to *both* carriers for a given germanide, accumulation carrier injection at the drain is increased. For a given device designed for a given off state current, the required sacrifice in drive current is necessarily larger for SFETs using pure germanium. It has thus been suggested that pure germanium is actually inferior to silicon regarding SFETs, due to this enhanced tunneling injection at the drain [20]. Such a conclusion is naturally valid for other semiconductors with relatively small bandgaps.

3.5. Controlling Ambipolarity

Due to the ambipolarity exhibited by SFETs of conventional design, it follows that tunneling injection at the drain places a lower limit on leakage current. Extending

this lower limit to provide a better off state requires a reduction in this tunneling injection, and therefore a reduction or elimination of ambipolarity for optimal CMOS functionality. As it turns out, in controlling this ambipolarity, the mechanisms of current modulation change.

One approach to controlling ambipolarity was taken by Lin *et al.*, whereby an asymmetric SFET, which was referred to as a field-induced drain (FID) device, was fabricated (Fig. 3.7) [21]. In their device, two gates were used to control current flow – a main-gate and a sub-gate. The main-gate acts as the gate in a conventional SFET, modulating the Schottky barrier at the source-body junction (although here it has no control over the drain-body junction). The sub-gate modulates the semiconductor region not covered by the main-gate (X_D in Fig. 3.7). In doing so, the sub-gate can create a thermal barrier within the body region to shut the device off, thus making the device unipolar in one direction or another, depending on the polarity of the applied biases.

Considering the case where the device in Fig. 3.7 is an n-body SFET, operating the device as a PFET would require negative gate and drain biases. Applying said biases to the main-gate, sub-gate, and drain would thus allow for p-channel operation. To switch the device off, the sub-gate bias is set to zero or some positive value. With the main-gate still at a negative bias, a thermal barrier within the body region arises, and the off state current is dramatically reduced (Fig. 3.8). This is because the off state current is now largely independent of the Schottky barriers at the source/drain regions. Achieving NFET-like operation simply requires reversing the bias polarities for the on and off states.

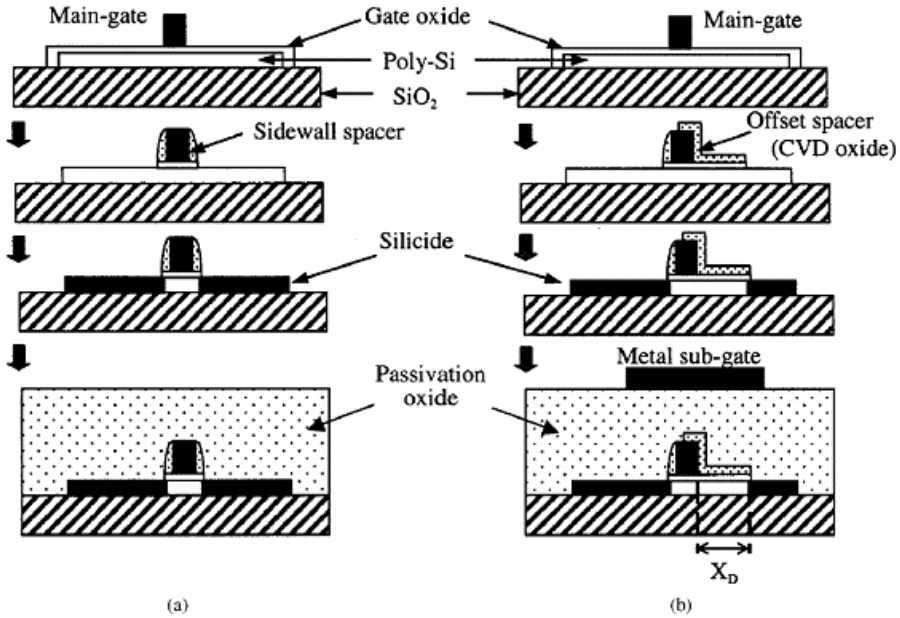

Fig. 3.7. Comparison of fabrication of conventional SFET (a) and field-induced drain (FID) SFET (b), adapted from [21]. The FID structure allows one to modulate the SFET behavior as explicitly NFET-like or PFET-like.

While the FID approach clearly shows strong control of ambipolar behavior with high I_{on}:I_{off} (~ 7 dec.), several disadvantages exist. First, the addition of a sub-gate results in a 4-terminal device for SOI substrates (bulk substrates would result in a 5-terminal device), which introduces additional interconnect routing challenges upon implementation into microelectronic circuitry. Second, the sub-gate is relatively far away from the portion of the body region under its control, and so high operating voltages are required (±50 V in the case of [21]). Certainly, the sub-gate could be placed closer to the device to reduce the voltage requirements, but it would not reduce the complexity of the device structure. Third, the asymmetry of the device demands that either the main-gate be made very small and/or the source-to-drain spacing be made larger to allow the sub-gate to "fit" within the device structure. Device/circuit density and scaling potential are thus compromised.

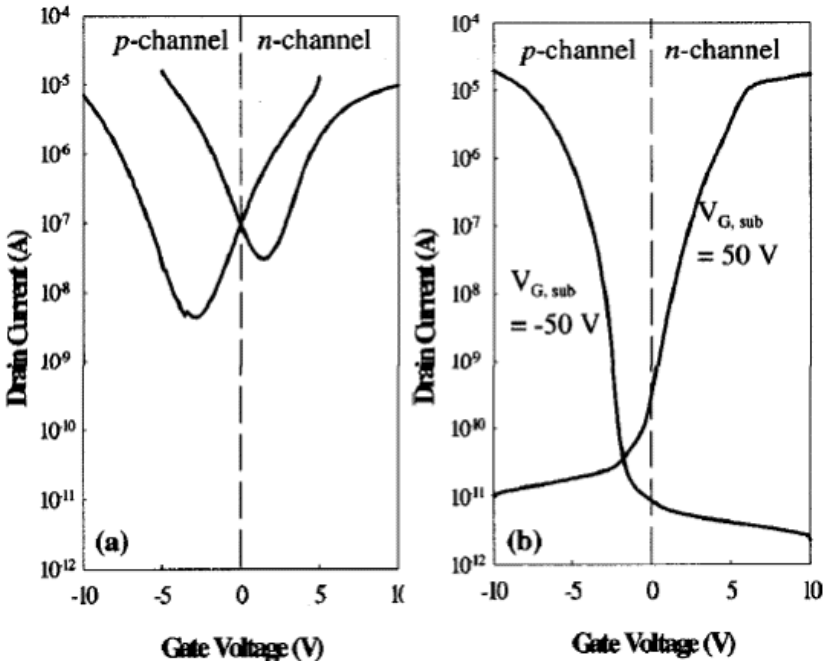

Fig. 3.8. Transfer curve comparison of a conventional ambipolar SFET (a) to the FID SFET (b), adapted from [21].

What is arguably the approach with the highest potential for integration into microelectronic circuitry is something called “bulk switching.” With this approach, the semiconductor near the source/drain regions is held at a different potential than the rest of the body region. This can be done electrostatically with separate gates [3] (not very practical) or chemically by utilizing halo regions [3], [4], [22]. Fig. 3.9 illustrates the band structure of a p-channel SFET using bulk switching. Much like what was done in [21], a thermal barrier created in the body region dominates current modulation, as the electric field at the Schottky barriers is large enough such that they can be considered ohmic. In the case of chemically formed halo regions, the higher dopant concentration prevents inversion within some voltage range, effectively making the device unipolar within this range. As the halo region and [lighter doped] body region are of the same

dopant type (i.e., n-type or p-type, although the actual species need not be the same), the device can be considered an accumulation mode device.

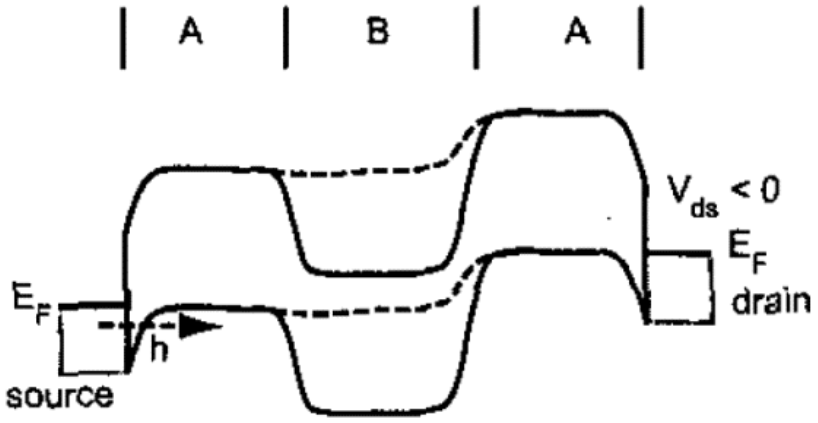

Fig. 3.9. Band structure of a bulk switching p-channel SFET or CNTFET, adapted from [3]. Regions A and B represent the halo and bulk regions, respectively. Region A can be formed with alternate gates (i.e., electrostatically) or chemically. The dotted lines represent the thermal barrier modulation that switches the device on and off.

To a first order, the bulk switching approach begs the question of whether such a device is truly an SFET, as Schottky barrier modulation is being replaced by thermal barrier modulation within the body region. However, the function of this halo region is to reduce the effect of the Schottky barrier on current injection. Since the source/drain regions are metallic, such a device is most certainly not a conventional MOSFET. The device is also not exactly an SFET in the conventional sense, as again the Schottky barrier is not modulated or being modulated very little by the gate; however, as the source/drain regions are silicided (or fully silicided in the case of SOI substrates), the actual characterization of such a device falls somewhere in the middle between an SFET and a conventional MOSFET – a metallic source/drain (MSD) MOSFET (the term “bulk switching SFET” is equally applicable). Regardless, it would seem that the MSD MOSFET is ultimately an evolution of the conventional MOSFET structure.

It comes with something of a sense of irony that the bulk switching approach seems the most promising, as the device engineer comes back full circle to the very challenge that they thought they would avoid by utilizing metallic source/drain regions with atomically abrupt junctions – dopant concentration gradients. As it turns out, however, the manner in which these halo regions are formed is quite different from the conventional approach, and this approach allows for extremely abrupt dopant concentration gradients compared to conventional MOSFETs. The process utilized is called implant to silicide, also known as implant through silicide (ITS). With ITS, the source/drain silicide regions are formed *before* the halo implantation. The higher atomic density of the metal silicide over silicon results in greater stopping power for a given implant energy, and that the dopants are implanted into the silicide eliminates any amorphization that would otherwise occur without ITS.

After ITS is performed, a relatively low temperature anneal, around 600°C [4], is performed to drive the dopants out of the silicide and into the adjacent silicon. Dopant diffusion in silicides is much greater than in silicon, and is attributed to grain boundary diffusion within the silicide [23]. Once the dopants reach the silicide-silicon interface, the diffusion slows down dramatically due to the lower dopant mobility in silicon. In [4], Monte Carlo simulation suggested that the dopant straggle distribution is 8 nm for their particular process. As their starting substrate had a background doping of $1x10^{15}$ cm^{-3}, and assuming a peak halo dopant concentration of $5x10^{19}$ cm^{-3}, the resulting junction abruptness between the body region and the halo region would be approximately 1.7 nm/dec. This value far exceeds the 2003 ITRS junction abruptness expectations at

the 65 nm node, and is about twice as abrupt as what is expected of circa 2005 front end process technology [24]. The resulting device structure is illustrated in Fig. 3.10.

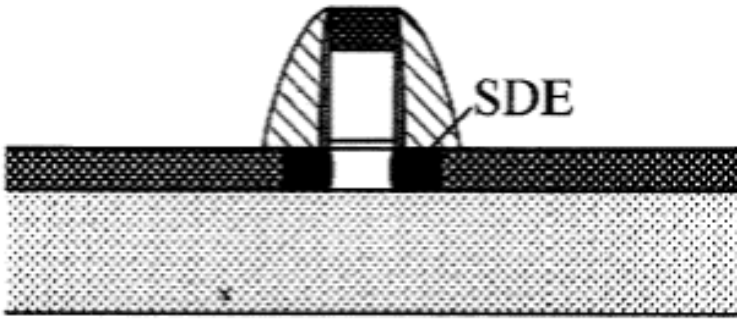

Fig. 3.10. Device structure of a bulk switching SFET, adapted from [4]. The source/drain extension (SDE) would be more appropriately referred to as a halo region, as it is of the same dopant type as the body region. Regardless, the relative simplicity of the structure relative to a modern conventional MOSFET is self-evident.

An additional advantage of utilizing bulk switching in SFETs is that the halo regions result in Schottky barrier lowering at the M-S junctions, thus reducing contact resistance. The extent of this barrier lowering is largely dependent on the dopant concentration in the halo regions, and is covered in more detail in Chapter 4. The utilization of bulk switching in SFETs does not come without its challenges, however. One such challenge is controlling the off state current. For the device in [4] (shown in Fig. 3.10), the halo and body regions are p-type, and the gate is a p+ poly gate. This results in the device being not entirely on or off when zero gate bias is applied, due to the relatively low to moderate channel resistance at the accumulated p-type surface (caused by the p+ poly gate). Switching to an n+ poly gate, or more preferably a low workfunction metal gate, should mitigate this problem considerably if not entirely, albeit at the expense of drive current.

However, in the essence of process simplicity, one might desire a single metal silicide for both the p-channel and n-channel devices, thus allowing for source/drain formation and fully-silicided (FUSI) gate formation at the same time for both devices. This places a preference toward midgap or near-midgap materials (such as NiSi). Such a midgap gate would indeed result in a better off state than, for example, the p+ poly gate for the device in [4]; however, whether the device is completely "off" (the flat region at positive V_G for V_d = -1 V in Fig. 3.11) depends at least in part on junction abruptness and/or channel length (modulating the channel length for a given junction abruptness modulates the percentage of the channel region occupied by the halo region). This is not to say that controlling off state current is difficult in bulk switching SFETs – it is simply to say that, for the simplest fabrication process, due attention must be paid to gate workfunctions and junction abruptness.

It is also noted that decreasing *EOT* will help the off state. For the 25 nm device in [4] (Fig. 3.11), the gate oxide thickness was 40 Å – very large for such a small device, where *EOT* ~ 10 Å would be more appropriate – and so the true potential of the bulk switching SFET in terms of both on state and off state current (as well as DIBL and subthreshold swing) was not realized. The results from [4] are therefore all the more impressive. Additionally, bulk switching SFETs are best utilized on SOI (or polysilicon-on-insulator – POI) substrates, as the relatively thin and isolated body regions substantially reduce sub-surface leakage, and gate workfunction engineering has a greater effect on the total off state current, due to the largely surface channel nature of such devices. It is noted that, for ultrathin body (UTB) SOI substrates (~ 10 nm and lower), the inversion channel consumes the entire body thickness and so the device is no longer a

"surface channel" device, although the point about gate workfunction engineering remains.

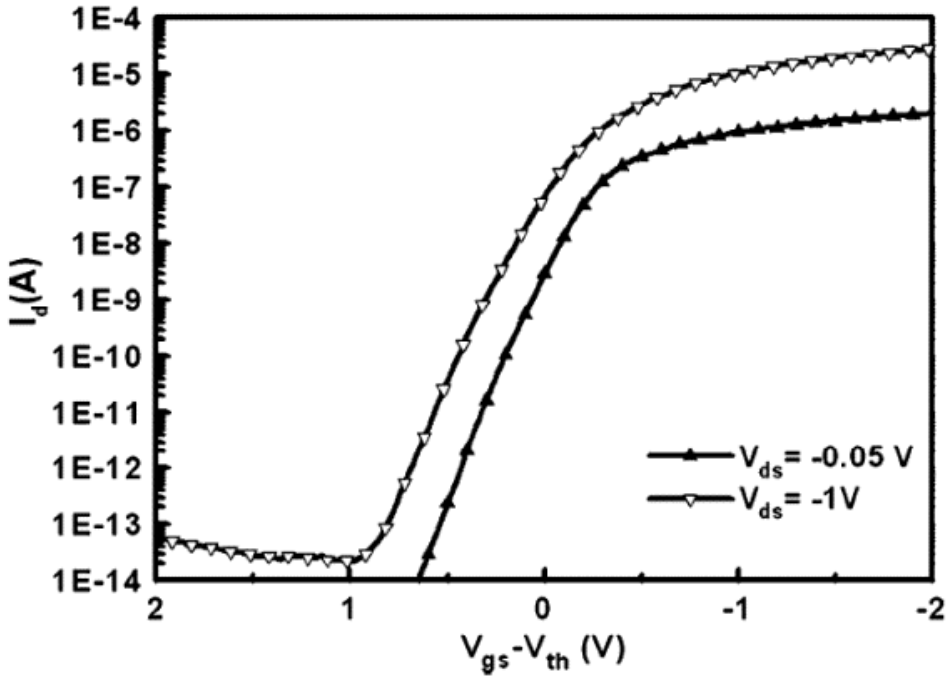

Fig. 3.11. Transfer characteristics of a modified Schottky barrier (MSB) FinFET (bulk switching SFET), adapted from [22].

Another challenge with bulk switching SFETs is band-to-band tunneling, as Fig. 3.12 illustrates. This is not so much an issue with silicon as it is with relatively low bandgap materials such as germanium or carbon [nanotubes]. In the example of an n-channel device, if the thermal barrier between the halo region and the body region is "large" (e.g., use of a high workfunction gate and low body doping or an undoped body), then energy states in the valence band in the body region can line up with energy states in the conduction band in the halo region, thus facilitating band-to-band tunneling. While this can happen in silicon-based devices, they have a larger bandgap than germanium

based or carbon nanotube based devices, which results in a wider tunneling barrier, and hence less tunneling.

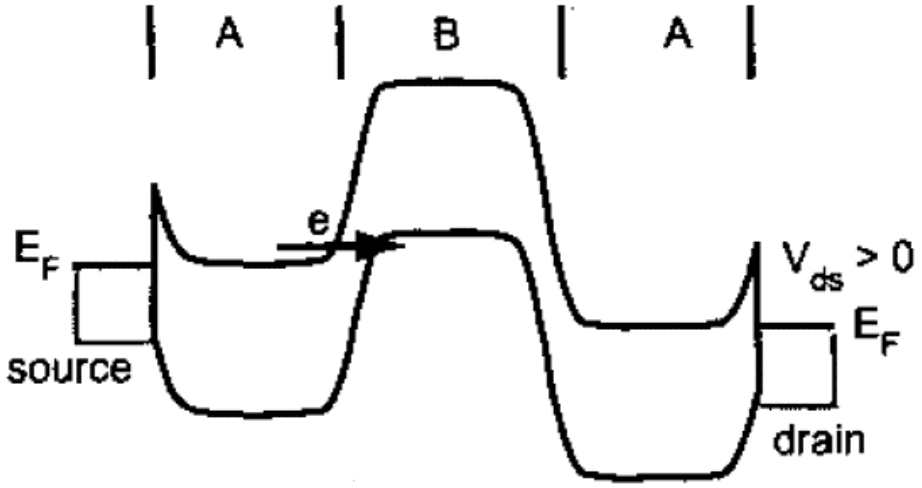

Fig. 3.12. Illustration of band-to-band tunneling leakage in an n-channel bulk switching SFET or CNTFET, adapted from [3].

In a test environment, for bulk switching devices with chemically doped halo regions, one can determine whether the leakage in the region of gate biases appropriate to induce band-to-band tunneling is actually said tunneling or if it is simply inversion of the halo regions (which would also result in ambipolar behavior) by plotting I_{DS} vs. V_{DS} in this region of gate biases. If band-to-band tunneling is the dominant leakage mechanism, then a negative differential resistance (NDR) region should be observed, because as V_{DS} increases, the valence-to-conduction band overlap at the drain decreases, thus increasing the tunnel barrier width, thus reducing band-to-band tunneling. The increase in tunnel barrier width is replaced by a decrease in thermal barrier height between the halo region at the drain and the lightly doped or undoped body region, and so I_{DS} increases again (this is not unlike the I-V characteristic of a tunnel diode). If inversion of the halo regions is

the dominant leakage mechanism in said range of gate biases, then the I_{DS} vs. V_{DS} curve would saturate with no NDR region, no matter how high V_{DS} is driven to.

For a given device using a given semiconductor, one can decrease the possibility of band-to-band tunneling leakage by creating another thermal barrier within the lightly doped/undoped body region. This is illustrated in Fig. 3.13, where there is an undoped portion of the body region directly at the center, surrounded by a lightly doped body region, surrounded by halo regions. The undoped portion at the center (region B') provides the same off state thermal barrier as before, but the lightly doped regions surrounding the undoped region result in a wider tunnel barrier. While this has been demonstrated using back-gated CNTFETs [3], a practical CMOS implementation would be more challenging, as patterning the undoped region requires a critical dimension that is considerably smaller than the gate length. With the appropriate device design, however, the "conventional" bulk switching SFET/CNTFET should prove sufficient for high performance CMOS at aggressive scales.

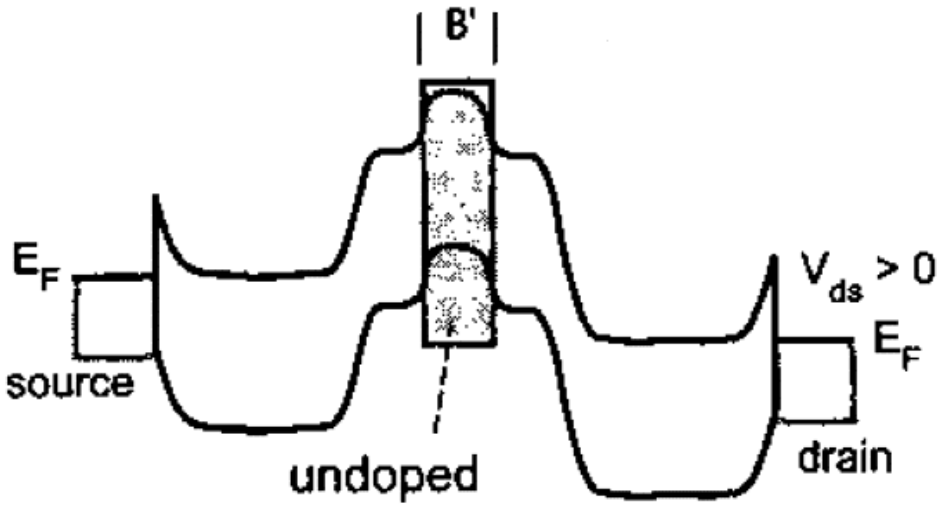

Fig. 3.13. Band diagram illustration of a bulk switching n-channel SFET or CNTFET with two thermal barriers within the body region, adapted from [3].

Chapter 3 References

[1] P. Avouris, R. Martel, S. Heinze, M. Radosavljevic, S. Wind, V. Derycke, J. Appenzeller, J. Terso, "The role of Schottky barriers on the behavior of carbon nanotube field-effect transistors," *Structural and Electronic Properties of Molecular Nanostructures*, 2002, pp. 508-512.

[2] P. Avouris, J. Appenzeller, R. Martel, S.J. Wind, "Carbon Nanotube Electronics," *Proc. of the IEEE*, Vol. 91, no. 11, 2003, pp. 1772-1784.

[3] Y.-M. Lin, J. Appenzeller, Ph. Avouris, "Novel Carbon Nanotube FET Design with Tunable Polarity," *IEEE Electron Devices Meeting*, 2004, pp. 687-690.

[4] B.Y. Tsui, C.P. Lin, "A Novel 25-nm Modified Schottky Barrier FinFET With High Performance," *IEEE Elec. Dev. Lett.*, Vol. 25, no. 6, 2004, pp. 430-432.

[5] L.E. Calvet, H. Luebben, M.A. Reed, C. Wang, J.P. Snyder, J.R. Tucker, "Subthreshold and scaling of PtSi Schottky Barrier MOSFETs," *Superlattices and Microstructures*, 2000, Vol. 28, no. 5/6, pp. 501-506.

[6] Y. Taur, T.H. Ning, "Fundamentals of Modern VLSI Devices," *Cambridge University Press*, 1998, pp. 128, 433.

[7] S. Zhu, H.Y. Yu, S.J. Whang, J.H. Chen, C. Shen, C. Zhu, S.J. Lee, M.F. Li, D.S.H. Chan, W.J. Yoo, A. Du, C.H. Tung, J. Singh, A. Chin, D.L. Kwong, "Schottky-Barrier S/D MOSFETs With High-K Gate Dielectrics and Metal-Gate Electrode," *IEEE Elec. Dev. Lett.*, Vol. 25, no. 5, 2004, pp. 268-270.

[8] J. Guo, M.S. Lundstrom, "A Computational Study of Thin-Body, Double-Gate, Schottky Barrier MOSFETs," *IEEE Trans. Elec. Dev.*, Vol. 49, no. 11, 2002, pp. 1897-1902.

[9] M. Jang, S. Lee, K. Park, "Erbium Silicided n-Type Schottky Barrier Tunnel Transistors for Nanometer Regime Applications," *IEEE Transactions on Nanotechnology*, Vol. 2, no. 4, 2003, pp. 205-209.

[10] J. Knoch, J. Appenzeller, "Impact of the channel thickness on the performance of Schottky barrier metal-oxide-semiconductor field-effect transistors," *App. Phys. Lett.*, Vol .81, no. 16, 2002, pp. 3082-3084.

[11] R.A. Vega, "On the Modeling and Design of Schottky Field Effect Transistors," *IEEE Trans. Elec. Dev.*, Vol. 53, no. 4, 2006, pp. 866-874.

[12] J. Larson, J. Snyder, "Schottky Barrier CMOS," pp. 1-12. Available: http://www.spinnakersemi.com/Documents/SBMOS_Customer_Technical_Whitepaper_Edition3.pdf

[13] J. R. Tucker, "Schottky Barrier MOSFETs for Silicon Nanoelectronics," *Proc. IEEE*, 1997.

[14] X. Liu, K. Juo, G. Du, L. Sun, J. Kang, R. Han, "N Channel SOI Schottky Barrier Tunneling Transistors," *Proc. IEEE,* 2001, pp. 562-565.

[15] D. Connelly, C. Faulkner, D.E. Grupp, J.S. Harris, "A New Route to Zero-Barrier Metal Source/Drain MOSFETs," *IEEE Trans. on Nanotechnology*, 2004, Vol. 3, no. 1, pp. 98-104.

[16] Q.T. Zhao, E. Rije, U. Bruer, St. Lenk, S. Mantl, "Tuning of Silicide Schottky Barrier Heights by Segregation of Sulfur Atoms," *Proc. IEEE,* 2004, pp. 456-459.

[17] D. Connelly, C. Faulkner, D.E. Grupp, "Optimizing Schottky S/D Offset for 25-nm Dual-Gate CMOS Performance," *IEEE Elec. Dev. Lett.*, Vol. 24, no. 6, 2003, pp. 411-413.

[18] H.M. Nayfeh, J.L. Hoyt, D.A. Antoniadis, "A Physically Based Analytical Model or the Threshold Voltage of Strained-Si n-MOSFETs," *IEEE Trans. Elec. Dev.*, Vol. 51, no. 12, 2004, pp. 2069-2072.

[19] G. Xia, H.M. Nayfeh, M.L. Lee, E.A. Fitzgerald, D. A. Antoniadis, D.H. Anjum, J. Li, R. Hull, N. Klymko, J.L. Hoyt, "Impact of Ion Implantation Damage and Thermal Budget on Mobility Enhancement in Strained-Si N-Channel MOSFETs," *IEEE Trans. Elec. Dev.*, Vol. 51, no. 12, 2004, pp. 2136-2144.

[20] S. Xiong, T.J. King, J. Bokor, "A Comparison Study of Symmetric Ultrathin-Body Double-Gate Devices With Metal Source/Drain and Doped Source Drain," *IEEE Trans. Elec. Dev.*, Vol. 52, no. 8, 2005, pp. 1859-1867.

[21] H.C. Lin, K.L. Yeh, T.Y. Huang, R.G. Huang, S.M. Sze, "Ambipolar Schottky-Barrier TFTs," *IEEE Trans. Elec. Dev.*, Vol. 49, no. 2, 2002, pp. 264-270.

[22] B.Y. Tsui, C.P. Lin, "Process and Characteristics of Modified Schottky Barrier (MSB) p-channel FinFETs," *IEEE Trans. Elec. Dev.*, Vol. 52, no. 11, 2005, pp. 2455-2462.

[23] K. Maex, M. Van Rossum, "Properties of Metal Silicides," *Short Run Press, Ltd.*, 1995, pp. 298- 306.

[24] International Technology Roadmap for Semiconductors (ITRS) (2003), Available: http://public.itrs.net

Chapter 4

Development of a Mathematical Model for SFETs

4.1 Model Approach

A relatively simple approach was taken to developing the mathematical model described in the following pages. First, an energy band model was developed to allow one to determine the valence band or conduction band behavior for a given device structure at varying gate and drain biases. These energy bands were then used to compute the total thermal barrier height to a given carrier in question, as well as the tunneling barrier width. Schottky barrier lowering (SBL) was incorporated into the model, although the "rounding off" at the top of the barrier was not accounted for – the abrupt characteristic of the ideal Schottky barrier was maintained, while simply decreasing its barrier height in accordance to the barrier lowering conditions. It is assumed in this model that the rounding off of the potential profile at the top of the Schottky barrier does not appreciably change the end result.

For the tunneling model, an Airy function transfer matrix approach was utilized, which is considered to be more accurate than the Wentzel-Kramers-Brillouin (WKB) approximation for very narrow barriers (i.e., high gate biases). The WKB model was also utilized in the essence of comparison. Total current injection was performed by taking a Riemann sum over a fine energy grid within a particular energy range, instead of performing an integral within this range (there is no mathematical justification for this – it was simply done to make the code writing easier). Model calculations were performed in MATLAB 7. It is noted that this model is a one-dimensional model, and only treats

current flowing at the surface of the device as a uniform sheet of charge. Subsurface current is either assumed or determined from empirical data, but in both cases is treated as independent of gate bias.

In addition, the use of fitting parameters, combined with the relative simplicity of the model, suggest that the model is not meant to predict specific behavior with high accuracy. Instead, it is best utilized in characterizing existing data to allow one to gain fundamental insight into the operation of a particular device structure, and to determine how specific changes to this structure may change the device performance. As it currently stands, this model does not include a universal mobility model for the channel region, and so ballistic transport is assumed. Furthermore, this model, in its current iteration, is specific to *single gate* devices, and assumes some arbitrary degree of gate overlap to the source/drain regions (which cannot be varied). Therefore, the effect of ultrathin body double gate devices, as well as the effects of overlap/underlap capacitance on high frequency operation, cannot be studied with this model. Again, this is a relatively simple model, developed for the sole purpose of gaining fundamental insight.

4.2 Energy Band Model

The energy band model used was adapted from the threshold voltage model of Liu *et al.* [1] and modified to result in a conduction or valence band profile. The expression for the modified energy band model is as follows:

$$\frac{-E}{q} = \mp[\psi_{slc} - V_{bi} + \phi_B + (V_{bi} \mp V_{DS} - \psi_{slc})\frac{\sinh\left(\frac{x}{l}\right)}{\sinh\left(\frac{L_{ch}}{l}\right)} + (V_{bi} - \psi_{slc})\frac{\sinh\left(\frac{L_{ch}-x}{l}\right)}{\sinh\left(\frac{L_{ch}}{l}\right)}] \quad (1)$$

$$l = \sqrt{\frac{\varepsilon_{si} t_{ox} W_{D\max}}{\eta \varepsilon_{ox}}} \tag{2}$$

$$\psi_{slc} = \mp V_{GS} + \phi_{ms} - \frac{qN_{sub} W_{D\max} t_{ox}}{\varepsilon_{ox} \varepsilon_0} + \frac{Q_{ox}^{'}}{C_{ox}^{'}} \tag{3}$$

where E is the carrier energy, V_{DS} is the drain-source voltage, L_{ch} is the channel length, x is the position across the channel, V_{bi} is the built-in voltage of the Schottky diode in equilibrium, ϕ_B is the Schottky barrier height (SBH), l is the characteristic length, defined in (2) [1], Ψ_{slc} is the long channel surface potential, defined in (3), ε_{si} and ε_{ox} are the relative dielectric constants of silicon and oxide, respectively, ε_0 is the vacuum permittivity, t_{ox} is the gate oxide thickness, W_{Dmax} is the maximum gate-induced depletion width into the body region, N_{sub} is the substrate doping level, q is the electron charge, V_{GS} is the gate-source voltage, ϕ_{ms} is the metal-semiconductor workfunction, $Q_{ox}^{'}$ is the oxide interface charge per unit area (set to zero for this work), $C_{ox}^{'}$ is the oxide capacitance per unit area, and η is a fitting parameter which has a dependence on L_{ch} and V_{DS}. Where there are varying ± signs, the top is for an n-type body region, while the bottom is for a p-type region (equations beyond this point assume an n-type body region). While this model was only noted in [1] to perform satisfactorily down to L_{ch} = 100 nm, it nevertheless provides a reasonable and relatively simple starting point for future work in modeling sub-100 nm SFETs.

With an energy band profile, it is now possible to find the electric field, ξ, at the source-body or drain-body junction at a given energy. This electric field is used in calculating the tunneling probability for a given carrier at a given energy, and is simply the derivative of (1). For electrons, the negative derivative is taken, while for holes, since

the energy bands must bend in the opposite direction to induce tunneling, the positive derivative is taken:

$$\xi = \mp\left(l \sinh\left(\frac{L_{ch}}{l} \right) \right)^{-1} [(V_{bi} - V_{DS} - \psi_{slc}) \cosh\left(\frac{W_B}{l} \right) - (V_{bi} - \psi_{slc}) \cosh\left(\frac{L_{ch} - W_B}{l} \right)] \quad (4)$$

where W_B is the tunnel barrier width at a given energy level. This is found by replacing x with W_B in (1) and solving for W_B. A quadratic in $exp(W_B/l)$ results, ultimately giving the expression shown in (5), where B_b, C_c, and D_d are defined in (6), (7), and (8), respectively.

$$W_B = l * \ln\left(\frac{B_b \pm \sqrt{B_b^2 - C_c D_d}}{C_c} \right) \quad (5)$$

$$B_b = \frac{\left(\frac{E}{q} + V_{bi} - \psi_{slc} - \phi_B \right) \sinh\left(\frac{L_{ch}}{l} \right)}{V_{bi} - \psi_{slc}} \quad (6)$$

$$C_c = 1 - \frac{V_{DS}}{V_{bi} - \psi_{slc}} - e^{-\frac{L_{ch}}{l}} \quad (7)$$

$$D_d = e^{\frac{L_{ch}}{l}} + \frac{V_{DS}}{V_{bi} - \psi_{slc}} - 1 \quad (8)$$

The use of the positive or negative square root term in (5) depends on the curvature of the energy band in question. Positive (convex – contact potential modulation) curvature requires the use of the positive square root, while negative (concave – Schottky barrier modulation) curvature requires the use of the negative square root. This is important to note; otherwise, the calculated barrier width becomes negative.

It was noted in [1] that l is proportional to $W_{Dmax}^{2/3}$, which implies that the fitting parameter η is also dependent on W_{Dmax}. Once η is found for a given technology,

however, it need not be changed if one intends to model the effect of modulating N_{sub}. Using the relationship between l and W_{Dmax}, a change in the characteristic length can be found with (9), where n_i is the intrinsic carrier concentration, N_{sub2} is the new substrate doping, l_2 is the new characteristic length, and N_{sub1} and l_1 are, respectively, the substrate doping and characteristic length of the original device from which η was found. This dependence between l and W_{Dmax}, however, was found in [1] empirically with devices having source/drain junction depths on the order of 0.2 μm to 0.35 μm, and so for the model presented here, the relationship is *assumed* to hold true for sub-100 nm devices.

$$l_2 = l_1 * \left[\frac{N_{sub1} * \ln\left(\frac{N_{sub2}}{n_i} \right)}{N_{sub2} * \ln\left(\frac{N_{sub1}}{n_i} \right)} \right]^{1/3} \tag{9}$$

It is important to note that the principal limitation of this energy band model is that it does not account for screening of the gate field by inversion or accumulation carrier charge. In other words, the sensitivity of the change in band bending on V_{GS} is only dependent upon l (which depends on η) and not the charge in the channel. For small values of l (large values of η), the change in surface potential is overestimated and may give misleading results. For non-negligible values of l, such as for the particular device structure discussed later where l is at least 1/3 of L_{ch} for the chosen values of η, this model seems to approximate the surface potential well enough to gain appropriate insight into the device behavior. A more accurate model would replace (3), which is directly proportional to V_{GS}, with a self-consistent solution, as well as account for carrier confinement at high gate fields. However, the complexity of the modeling code would increase due to the lack of a closed form solution to Ψ_{slc} for a given V_{GS}.

4.3 Tunneling Models

There are three approaches in particular that are used to model tunneling current in SFETs – Airy functions in a transfer matrix [2], [3], the Wentzel-Kramers-Brillouin (WKB) approximation [4], [5], and a self-consistent solution of the Poisson and Schrödinger equations, usually in one dimension [6] or two dimensions [7], [8]. The WKB approximation is computationally simpler than the Airy function approach, but it is only accurate for slowly varying potentials (small electric fields) [9]. The self-consistent solutions performed in [6]-[8] are very interesting, but it was not made clear whether and/or how SBL was included. Also, the mathematical complexity, and hence computational cost, can be relatively high. The Airy function approach reaches a "middle ground," whereby high accuracy can be achieved with a moderate mathematical complexity and computational cost. The Airy function approach and the WKB approximation will be explored in this discussion.

The tunneling probability from the WKB approximation, modified to include SBL and assuming a triangular potential profile, can be expressed as:

$$T_{TL\_WKB} = \exp\left(\frac{-4\sqrt{2qm^{*}m_0}\left(\xi \cdot W_B - \Delta\phi_B\right)^{3/2}}{3\hbar \cdot \xi} \right) \tag{10}$$

where m^* is the effective carrier mass ($0.26m_0$ for electrons and $0.36m_0$ for holes, where m_0 is the electron rest mass), $\hbar$ is Planck's constant (in units of J·s), $\Delta\phi_B$ is the SBL term, and ξ is the electric field from (4), but in units of V/m (likewise, the W_B term would be in meters, whereas in (5) it is in units of cm).

Normally, the barrier height term used in the WKB approximation is $(\phi_B - \Delta\phi_B - E)^{3/2}$; however, this approach does not account for the barrier width – there is only a

barrier height and lateral field dependence. At gate biases below the source-body flatband voltage, V_{sbfb}, a conventional thermal barrier produced by the dopant-induced and/or gate workfunction-induced band bending within the body region is modulated [10]. Beyond V_{sbfb}, the Schottky barrier is modulated; however, even for $|V_{GS}| < |V_{sbfb}|$ there exists a tunneling barrier to inversion carriers at some energies, an example of which is shown in Fig. 4.1. This barrier is much wider than the Schottky barrier under inversion, and so for the WKB model, the mathematical effect is a change in the barrier height. This is accounted for in the $\xi \cdot W_B$ term in (10), in which the E and ϕ_B dependencies are embedded. Without the inclusion of W_B dependence in the WKB model, the calculated tunneling probabilities for both cases in Fig. 4.1, where the electric fields are roughly equal, would be roughly the same.

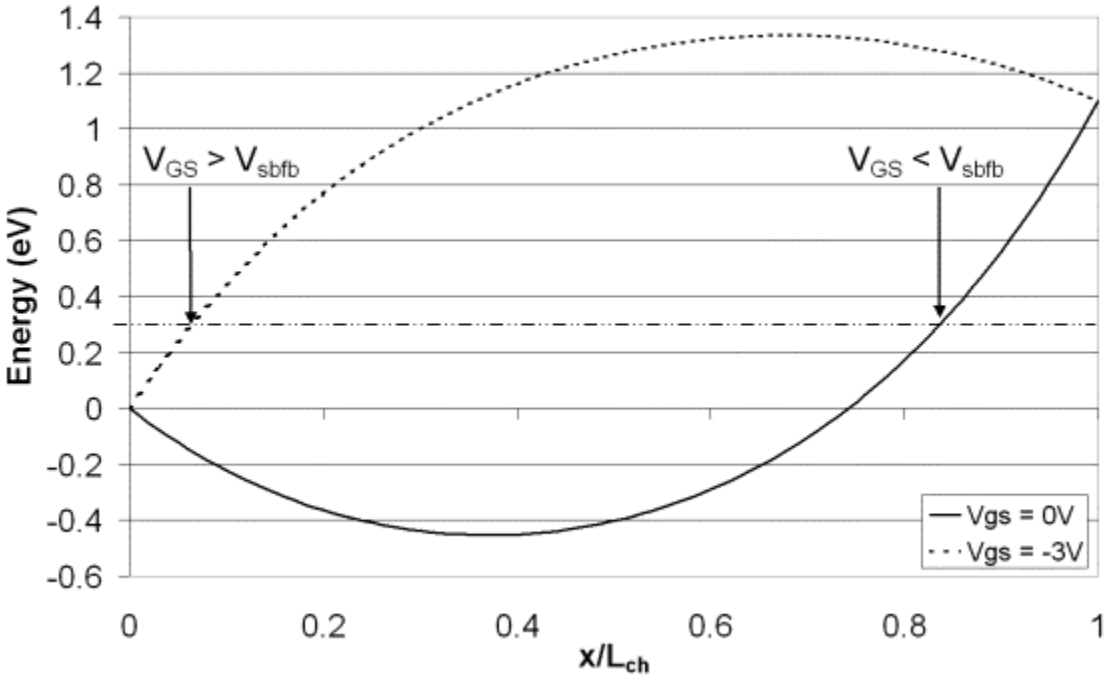

Fig. 4.1. Example valence band profiles of a p-channel SFET. Energies are taken relative to the hole SBH.

The Airy function tunneling model used was adapted from the work of Brennan and Summers [9], which used Airy functions and transfer matrices in a multiquantum well structure. This model, although more complicated, is considered to be much more accurate than the Wentzel-Kramers-Brillouin (WKB) approximation commonly used when modeling tunneling current. As a Schottky barrier presents only one quantum mechanical tunnel barrier, only one transfer matrix is used. Throughout this chapter, unless otherwise noted, tunneling current modeling will be performed with this Airy function model. The expression for tunneling probability is as follows [9]:

$$T_{TL} = \left(\frac{4k}{k'}\right) * \left[\left(A + \left[\frac{k'}{kD}\right]\right)^2 + \left(\frac{C}{k} - k'B\right)^2\right]^{-1} \tag{11}$$

where T_{TL} is the tunneling probability, A, B, C, and D are elements of the $S(0,W_B)$ transfer matrix, shown in (13)-(16), and k is expressed as:

$$k = q\sqrt{\frac{2m^* m_0 E}{\hbar^2}} \tag{12}$$

where m^* is the effective carrier mass (0.26 for electrons and 0.36 for holes in silicon), m_0 is the electron rest mass, and $\hbar$ is Planck's constant (in units of J·s) In the application to SFETs, k and k' are equal (or rather, using k' with $E = E' + qV_{DS}$, where E' is the Fermi energy at the source terminal, made no noticeable difference). The elements of the $S(0,W_B)$ matrix are:

$$\Gamma * A = A_i(\rho(x=0))B_i'(\rho(x=W_B)) - B_i(\rho(x=0))A_i'(\rho(x=W_B)) \tag{13}$$

$$\Gamma * B = B_i(\rho(x=0))A_i(\rho(x=W_B)) - A_i(\rho(x=0))B_i(\rho(x=W_B)) \tag{14}$$

$$\Gamma * C = A_i'(\rho(x=0))B_i'(\rho(x=W_B)) - B_i'(\rho(x=0))A_i'(\rho(x=W_B)) \tag{15}$$

$$\Gamma * D = B_i'(\rho(x=0))A_i(\rho(x=W_B)) - A_i'(\rho(x=0))B_i(\rho(x=W_B)) \tag{16}$$

$$\Gamma = A_i\left(\rho\left(x=0\right)\right)B_i'\left(\rho\left(x=0\right)\right) - B_i\left(\rho\left(x=0\right)\right)A_i'\left(\rho\left(x=0\right)\right) \tag{17}$$

where $A_i(\rho)$, $A_i'(\rho)$, $B_i(\rho)$, and $B_i'(\rho)$ are Airy functions of the first and second kind and their derivatives. Γ is a common term to (13) – (16) that results from the matrix algebra to find the $S(0,W_B)$ matrix. The expression for the spatial parameter $\rho(x)$ is given as:

$$\rho\left(x\right) = \left(x + \frac{\phi_B - \Delta\phi_B - \frac{E}{q}}{\xi} \right)\left(\frac{2qm^* m_0 \xi}{\hbar^2} \right)^{\frac{1}{3}} \tag{18}$$

where x and ξ are in units of m and V/m, respectively, E is taken from (1), and $\Delta\phi_B$ is the SBL induced by the lateral field, expressed as:

$$\Delta\phi_B = \sqrt{\frac{q\alpha|\xi|}{4\pi\varepsilon_0\varepsilon_{si}}} \tag{19}$$

where α is a fitting parameter. Classical theory sets α to 1 [12]; however, fitting this model to empirical data (shown later) suggests that this value varies and depends on V_{DS} and L_{ch}. In terms of physical meaning, α may be related to a dependence of SBL on Heine tail decay, as suggested in [13], and so hints at barrier lowering mechanisms beyond those induced by the lateral field alone [13], [14]. It may also account for, to some extent, the exclusion of gate field screening at large surface potentials. Although in this discussion the determination of α is somewhat arbitrary, the resulting extent of barrier lowering to inversion carriers of ~0.1 – 0.15 eV (shown later) to achieve a data fit is similar to that in [15]. Also, as mentioned previously, this model assumes an abrupt Schottky barrier under all electric field conditions, and so the fitting parameter may also account for the difference between the modeled potential profile and an actual potential profile, for which the top of the Schottky barrier is rounded off. It should be noted that

(13)-(16) are set up to model tunneling at the source of an SFET. For tunneling at the drain (i.e., accumulation carrier leakage), the $x = W_B$ term is replaced with $x = L_{ch} - W_B$.

4.4 Contact Potential

As discussed in Chapter 3, part of the subthreshold region of operation in SFETs involves modulation of the contact potential, ϕ_c, which is a thermal barrier presented to carriers beyond the SBH. For inversion carriers, this thermal barrier is induced by body dopants and/or the gate workfunction, either of which will result in band bending (and therefore a contact potential) in the body region. The contact potential is greater than zero for carriers at the source when ξ in (4) is negative. When this condition is met, finding the contact potential requires finding the position x_{max} where ξ is zero and then solving for the energy, E, in that position. To do this, W_B in (4) is replaced with x_{max}, which is then solved by setting ξ to zero. This position is plugged into (1), in which ϕ_B is set to zero. The solution for x_{max} is shown in (20). For long channel behavior ($L_{ch} >> l$ and/or small V_{DS}), (20) reduces to $L_{ch}/2$. Once V_{GS} reaches the source-body flatband voltage, V_{sbfb}, ϕ_c becomes zero and now the Schottky barrier is modulated. For SFETs fabricated on bulk substrates, ϕ_c is of particular importance, as it allows for an exponential control of subsurface leakage.

$$x_{\max} = \frac{l}{2} * \ln\left(\frac{e^{\frac{L_{ch}}{l}} - 1}{\left(\frac{V_{bi} - V_{DS} - \psi_{slc}}{V_{bi} - \psi_{slc}} \right) - e^{\frac{-L_{ch}}{l}}} \right) \tag{20}$$

It is also noted that, for long channel behavior, $\phi_c = V_{bi}$. As the channel length is reduced and the source/drain depletion regions begin to overlap, ϕ_c starts to decrease and

eventually reaches zero. The same effect occurs as V_{DS} is increased (DIBL). The extent of short channel behavior, then, can also be monitored by observing the change in V_{sbfb} with L_{ch} for $V_{DS} = V_{DD}$ (the power supply voltage). One can thus determine when punchthrough occurs by finding where $V_{sbfb} = 0$ V.

4.5 Obtaining the Total Current

By understanding the carrier transmission (tunneling or thermal) at a given energy, the total transmission can be found by integrating over an energy range. For a 1-dimensional system, and under the assumption of ballistic transport, the tunneling current density and thermal current density at the surface are found as shown in (21) and (22), where A_R is the effective Richardson's constant (112 A/cm^2·K^2 for electrons, 32 for holes in silicon), T is the temperature, k_B is Boltzmann's constant, ϕ_{Binv} and ϕ_{Bacc} are the inversion mode and accumulation mode SBHs (integration takes place over the energy gap of the semiconductor), respectively, and f_s and f_d are the Fermi-Dirac (F-D) distributions at the source and drain, respectively. As the F-D distribution shows the probability of a carrier existing at a given energy, and T_{TL} shows the probability of a carrier tunneling through the Schottky barrier in question at a given energy (in the case of (22), T_{TL} is 1), the integrands in (21) and (22) effectively model the probability of a carrier existing in the channel at a given energy. Multiplying this probability by A_R*T/k_B results in the density of injected current through and over the Schottky barrier in question.

$$J_{tun} = \frac{A_R T}{k_B} \int_{-q\phi_{Binv}}^{q\phi_{Bacc}} T_{TL}(E)\left[f_s(E) - f_d(E)\right] dE \qquad (21)$$

$$J_{therm} = \frac{A_R T}{k_B} \int_{q(\phi_B + \phi_c - \Delta\phi_B)}^{\infty} \left[f_s\left(E \pm q\Delta\phi_B\right) - f_d\left(E \pm q\Delta\phi_B\right) \right] dE \qquad (22)$$

Model calculations were performed in MATLAB 7, although instead of performing the integrals in (21) and (22), Riemann sums over a fine energy grid (ΔE = 0.01 to 0.001 eV) were performed. In the case of the thermal current calculation, the upper limit of the "integral" is not taken to infinity – a value between 1 and 3 (depending on V_{DS} and whether one is modeling current at the source or at the drain) is equally useful, as the F-D distribution at increasing energies becomes very small very quickly at room temperature.

Since $\Delta\phi_B$ causes a shift in both the electron and hole barrier heights, for the thermal current only, this is equivalent to an energy shift of the entire bandgap relative to the Fermi level in the metallic source/drain regions. This can be expressed mathematically by shifting the F-D distributions in the source/drain regions by $\Delta\phi_B$ (positive for electrons, negative for holes). In [3], the integration limits for (21) were not shifted by $\Delta\phi_B$ in order to maintain relative computational simplicity and to achieve smoother curves with a larger energy grid (less computational time). However, this resulted in the inclusion of part of the F-D distribution above ϕ_B - $\Delta\phi_B$, thus resulting in an overestimation of the tunneling current, which is quantified in Fig. 4.5 in Section 4.7 for both the WKB model and the Airy function model.

In some approaches [4], [5], the F-D distribution is replaced with a Maxwell-Boltzmann (M-B) distribution when modeling thermal current, as the distribution tends to be small above the Schottky barrier, in which case the F-D and M-B distributions can be used interchangeably. This also simplifies the integral in (22). Such an approach is not

taken here, however, as $\Delta\phi_B$ shifts the position of ϕ_B relative to the Fermi level, effectively changing the distribution above the Schottky barrier. Using an M-B distribution in such a case can result in a drastic overestimation of the thermal current for small barrier heights and high lateral fields. That is, as SBL increases, the difference between the solutions to the M-B and F-D distributions grows considerably.

The integrals in (21) and (22) give current density in units of A/cm^2. To acquire units of μA/μm, the channel is assumed to exist at the surface uniformly with a depth of 5.56 nm, and so (21) and (22) are multiplied by a constant of 5.56x10^{-5} μA·cm/A. This channel depth is a linear extrapolation from the observation that, in SFETs, 90% of the current is injected within the first 5 nm below the surface [2].

4.6 Comparison to Data

Model calculations were compared with data (extracted by hand) from the sub-30 nm and sub-80 nm p-channel SFETs from [16], shown in Figs. 4.2 and 4.3, respectively. These particular devices were chosen because of the information available regarding the design details. For both devices, the source/drain regions are PtSi (ϕ_B ~ 0.87 eV to electrons, ~ 0.25 eV to holes), N_{sub} = 1x10^{18} cm^{-3}, t_{ox} = 18 Å, n+ poly gate. L_{ch} was set to 25 nm and 75 nm for the sub-30 nm and sub-80 nm devices, respectively. V_{GS} was driven out to 3 V because of the relatively thick t_{ox}, for which 3 V puts the oxide field at about 7 MV/cm [16], which is typical for current devices. For each plot, α and η were adjusted separately for electrons and holes. The "model" for sub-surface leakage was simply a chosen current density to result in the appropriate minimum for the total current density. A more rigorous prediction of sub-surface leakage would require a 2-

dimensional model.

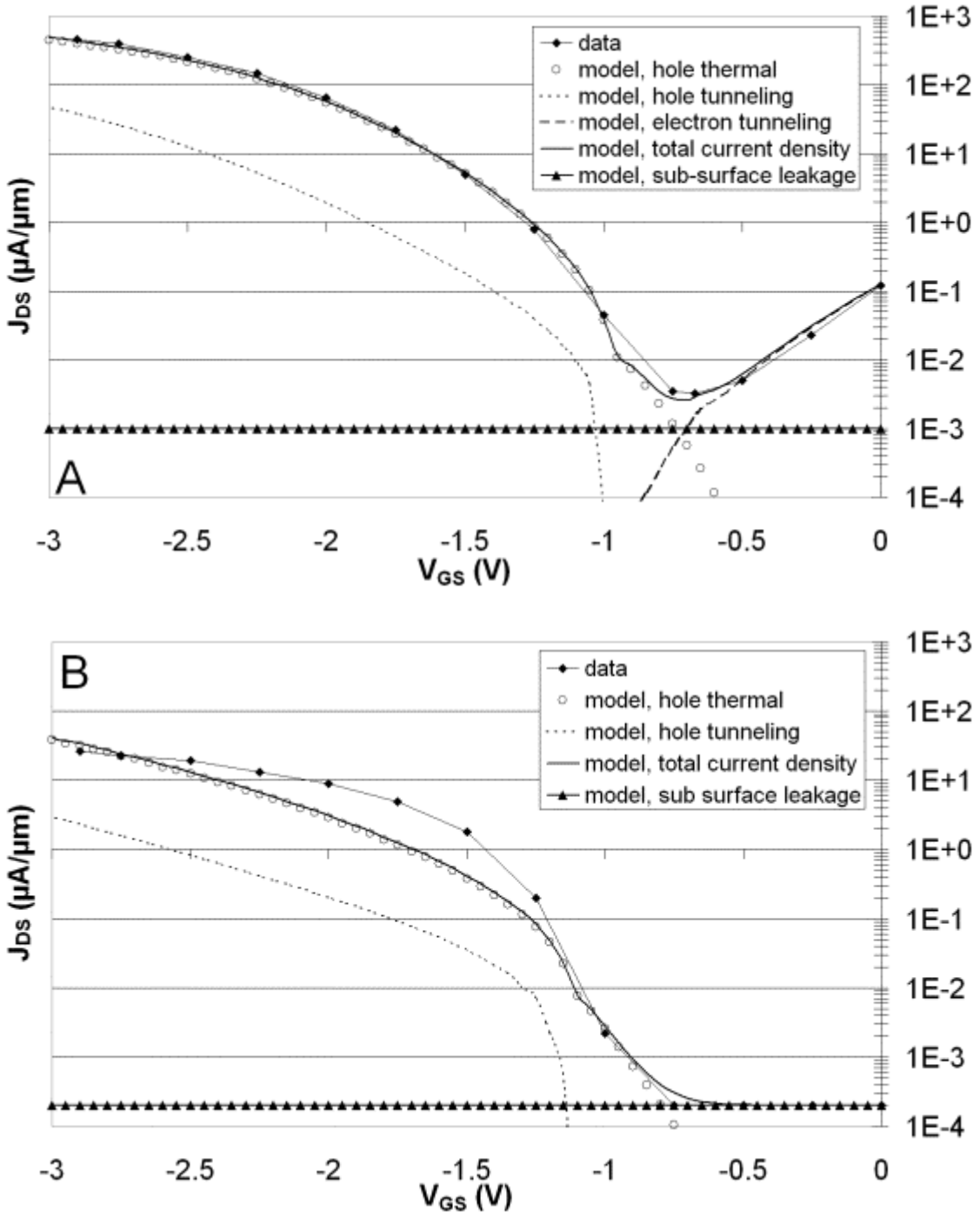

Fig. 4.2. Transfer characteristics for 25 nm p-channel SFET. In (a), V_{DS} = -1.1V, α and η are 1.1 and 2.6, respectively, for holes, and 0.1 and 1.1, respectively, for electrons. In (b), V_{DS} = -0.1 V, α and η are 0.9 and 1.2, respectively, for holes, while electron injection was not included. As both figures show, the tunneling contribution in the on state is considerably lower than the thermal contribution.

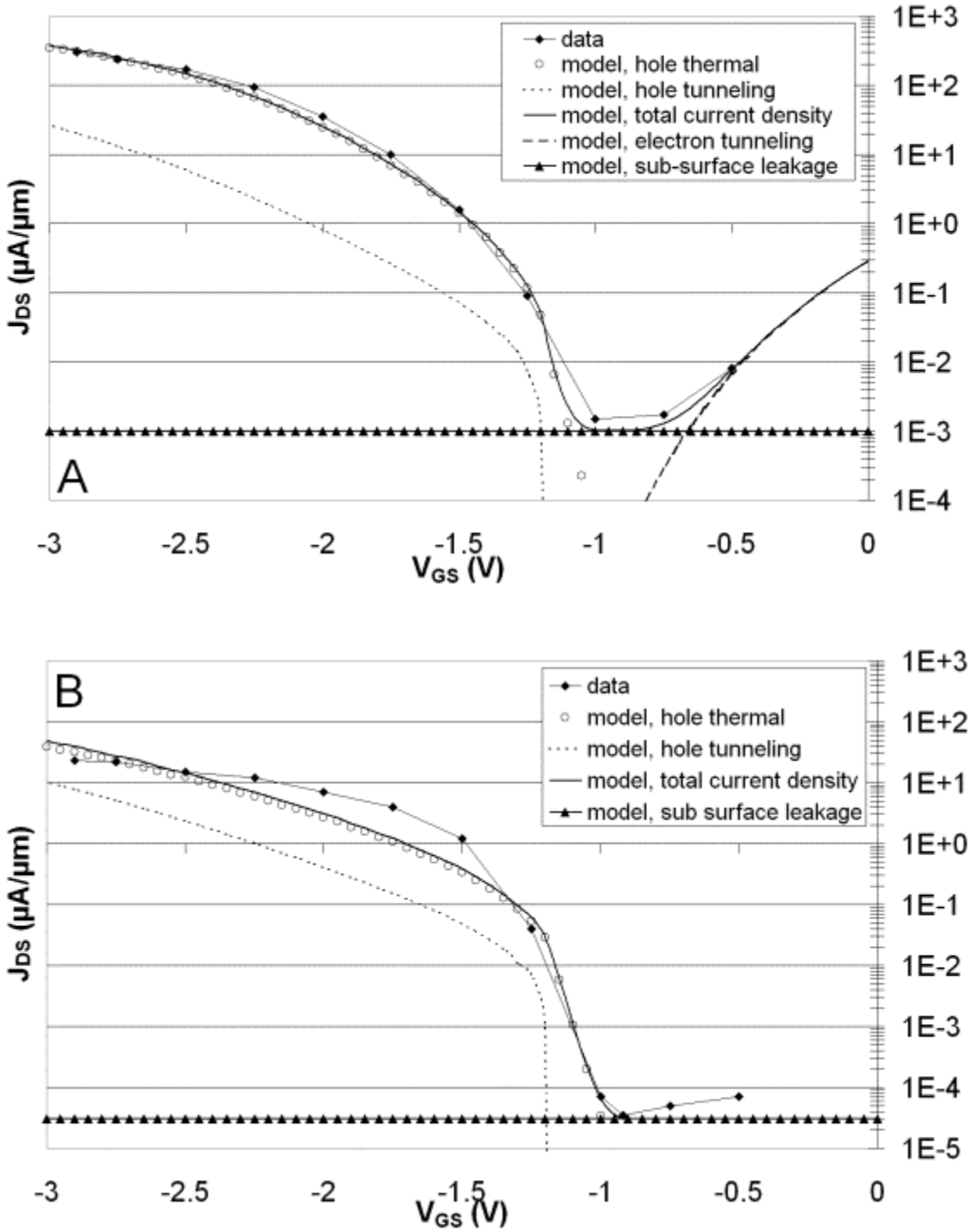

Fig. 4.3. Transfer characteristics for 75 nm p-channel SFET. In (a), V_{DS} = -1.1 V, α and η are 1.1 and 1.8, respectively, for holes, and 0.1 and 0.9, respectively, for electrons. In (b), V_{DS} = -0.1 V, α and η are 1.1 and 1.2, respectively, for holes, while electron injection was not included. Much like Fig. 4.2, the tunneling contribution in the on state is considerably lower than the thermal contribution.

While a very good fit to data can be achieved with this model for the saturation mode of operation (Figs. 4.2a and 4.3a), it is clear that the linear mode of operation (Figs. 4.2b and 4.3b) does not allow for the same type of fit. Most likely, this has to do with the aforementioned exclusion of a universal mobility model, as well as screening of the gate field by charge in the channel. For low V_{DS} and moderate to high V_{GS}, the vertical field is large enough to result in mobility degradation due to surface scattering at the semiconductor-gate dielectric interface. At high V_{DS}, surface scattering is very small if not negligible, and for small enough devices (sub-100 nm), carrier transport can be ballistic or near ballistic (depending on temperature, body dopant concentration, etc.). Thus, the channel mobility has a stronger dependence on gate bias for small V_{DS} than for high V_{DS}, which is an effect that the fitting parameters α and η do not seem to be able to account for in terms of achieving a fit to data.

Including a more accurate solution to the Poisson equation (i.e., gate field screening) and a universal mobility model would naturally change the required α and η values to achieve a data fit. More specifically, η may have to be increased, while α would be changed to account for any discrepancy that the channel mobility does not cover for $V_{GS} > V_{sbfb}$. This lends weight to the idea that the ballistic transport assumption is not entirely valid, even in the saturation mode transfer curves, although the extent of this discrepancy has been left to future modeling studies. The important point that this uncovers is that modeling SBL in SFETs with appropriate rigor may not be as simple as using (19) with α set to 1, and that in this device structure, the Schottky barrier might not be the limiting factor. While the Schottky barrier is a limiting factor to drive current in some device structures, *all device details must be considered* (i.e., channel mobility

reduction due to increased body doping or ultrathin body regions, *EOT*, etc.).

As this model suggests, regardless of the values chosen for α and η, tunneling current does not dominate current flow for these particular devices when the Airy function method is used, as it was in [3]. In the case of Figs. 4.2 and 4.3, the hole tunneling contribution to the total on state current is, at best, on the order of a few percent (Fig. 4.4), and hole thermal current dominates regardless of whether ϕ_c or ϕ_B is being modulated. Increases in the total current for V_{GS} beyond V_{sbfb} are therefore due primarily to the increase in *thermal current* from SBL and *not* the increase in tunneling current from Schottky barrier narrowing. While the tunneling transfer curve follows a similar slope to the total current density for $V_{GS} > V_{sbfb}$, this can be misleading or mistaken to dominate the total current in this region.

Looking further at Fig. 4.4, the hole tunneling percent of the total current does not even break 10 % for the particular device structure modeled. Recalling that the tunneling current modeled in [3] was an overestimation, the actual maximum contribution for this device is closer to 5 %, as will be shown in Section 4.7. As expected, the 25 nm device, which is of the same structure (i.e., body doping, gate dielectric thickness, etc.), has a lower $|V_{sbfb}|$ than the 75 nm device by about 200 mV. This suggests a greater SCE for the 25 nm device, which is also reflected in the results for hole thermal barrier height versus V_{GS} in Fig. 4.4. For the 25 nm device, this barrier height is always smaller, especially below V_{sbfb}. The somewhat parabolic behavior for the hole tunneling percent of total current for V_{GS} beyond V_{sbfb} has to do with carrier action at "low" and "high" lateral fields at the source-body junction. As V_{GS} increases beyond V_{sbfb}, tunneling current at the top of the barrier is replaced by thermal current at that energy as $\Delta\phi_B$ drops ϕ_B below its

previous value. At high V_{GS}, the lateral field is very strong (W_B is very small), and so the hole tunneling contribution increases, but never surpasses the thermal current, as the lateral field also causes barrier lowering.

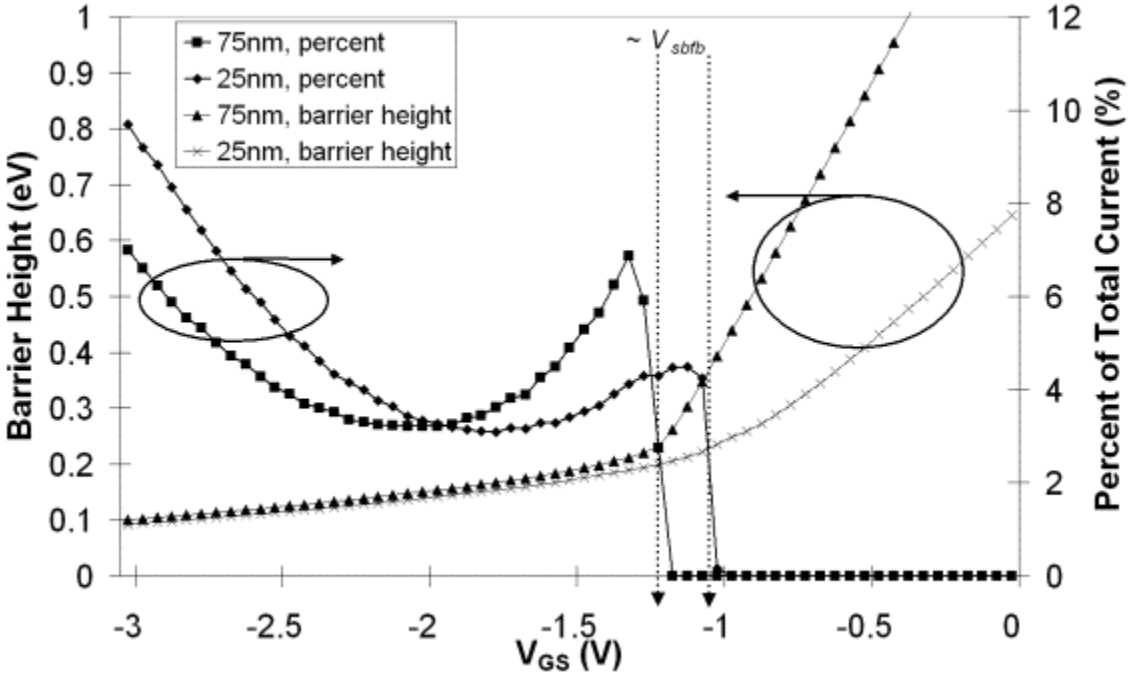

Fig. 4.4. Hole thermal barrier height and tunneling percent of total current versus gate bias for the 25 nm and 75 nm p-channel SFETs in this discussion. Although the SBL term drops ϕ_B to ~ 0.1 eV, the modeled tunneling current contribution in both devices does not exceed 10 % of the total current.

This effect of dominating thermal current only occurs within a range of barrier heights, though, as for a large barrier height (such as PtSi to electrons), the thermal current is very small. In such a case, and under very high lateral fields, tunneling current dominates over thermal current. This is shown in Figs. 4.2a and 4.3a at low $|V_{GS}|$, where electron tunneling injection at the drain due to the high body doping is the dominant leakage mechanism. Although data for electron thermal current is not included in Figs. 4.2 and 4.3, it was calculated to be less than $1x10^{-10}$ μA/μm, and so was treated as negligible.

4.7 Comparison of Tunneling Models

Regarding the case of "shifted" and "non-shifted" integration limits for (21) mentioned earlier, Fig. 4.5 shows the disparity between these two approaches for both the Airy function and the WKB model. In the particular case of the device modeled, V_{sbfb} is around –1 V, and so data from –1 V to –3 V is the data of interest here. As Fig. 4.5 shows, when the integration limits are not shifted, the overestimation in tunneling current is on the order of 160-220 %, regardless of whether the Airy function model or the WKB model is used. Technically speaking, in this analysis, the integration limits were not exactly shifted, but instead the part of the F-D distribution above ϕ_B - $\Delta\phi_B$ as a function of V_{GS} was removed. This results in some underestimation of the tunneling current, as tunneling at the other end of the energy gap deep below the SBH becomes excluded. The actual tunneling current prediction from either model, therefore, would be somewhere in between the overestimated and underestimated cases, but much closer to the underestimated case, as tunneling deep below the SBH (where W_B is larger) provides a smaller contribution to the total tunneling current. The differing behavior between the WKB and Airy function models in Fig. 4.5 can be traced back to the exclusion of tunneling deep below the SBH. As both models provide a different tunneling probability as a function of energy, the contribution of tunneling deep below the SBH is naturally different between the two models, thus resulting in at least part of the observed difference in Fig. 4.5. For the rest of this section, the "underestimated" case (i.e., shifted integration limits) is used unless otherwise noted.

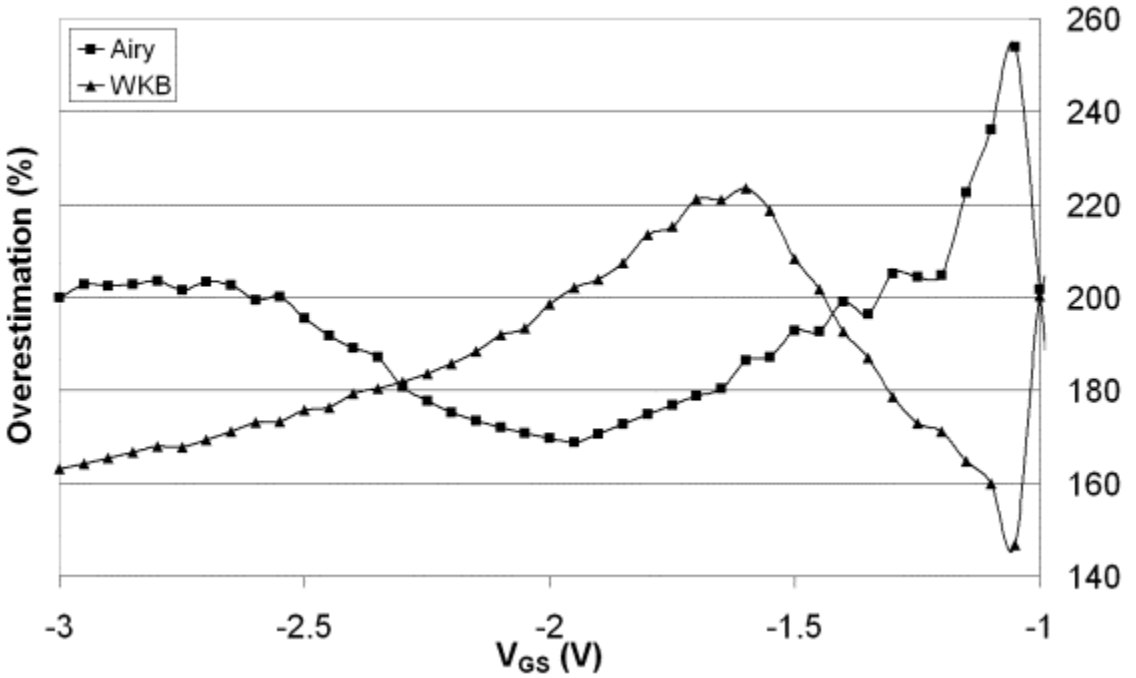

Fig. 4.5. Percent overestimation of tunneling current versus gate bias for the Airy function and WKB tunneling models when SBL is included. This overestimation is *not* an overestimation of the actual tunneling current, but of the modeling discrepancy between the cases of shifted and non-shifted integration limits. Smoother curves can be achieved with a finer grid structure, but this imposes a tradeoff between curve smoothness and computational time.

There is an issue of computational requirements when one considers the inclusion or exclusion of shifted integration limits for the tunneling current calculations. The code required to "shift" the integration limits (or rather, to remove the F-D distribution above ϕ_B - $\Delta\phi_B$, as mentioned earlier) is not terribly complicated; however, the effect is a change in the smoothness of the tunneling transfer curve, as some portions of the energy grid are not used throughout the entire V_{GS} range. Much like the thermal current calculation, very fine energy grid spacings of 0.001 eV or lower are required to achieve a smooth curve, and so the computational time is increased. In the case of non-shifted integration limits, the curve smoothness is largely independent of the size of the grid spacing – only the accuracy is dependent on the grid spacing, and so $\Delta E \sim 0.01$ eV gives very nice results.

However, as not shifting the integration limits is not as accurate a representation of the device behavior, the resultant tunneling current calculations do not exactly represent the "actual" tunneling current one should expect, as Fig. 4.5 illustrates. A "cheap" approach out of this predicament would be to simply divide the results of the non-shifted integration limits by 2, since Fig. 4.5 shows that the aforementioned overestimation averages out to somewhere around 200 %. None of the figures in this chapter use such an approach, though, as it does not have a rigorous physical foundation.

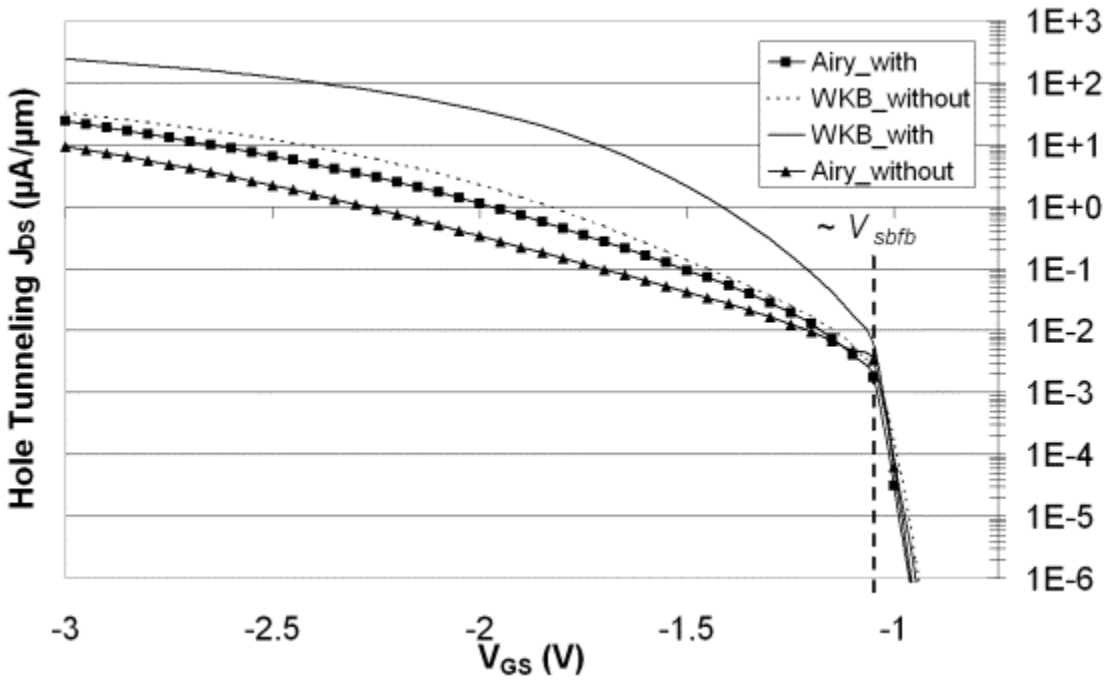

Fig. 4.6. Hole tunneling vs. V_{GS} for the 25 nm p-channel SFET using the Airy function and WKB models, with and without SBL. In both cases, the WKB model predicts greater tunneling current than the Airy function model.

Fig. 4.6 contains the model results for hole tunneling current density in the 25 nm p-channel SFET for both the Airy function model and the WKB model, with and without SBL. For $V_{GS} < V_{sbfb}$, all four cases give almost the same result. This is not surprising, as in this region, the contact potential is still being modulated, and so the tunneling barrier is

still very wide (Fig. 4.1). Beyond V_{sbfb}, the WKB model estimates a higher tunneling current than the Airy function model, with and without SBL. Although the WKB model without barrier lowering gives a similar result to the Airy function model with barrier lowering, the exclusion of barrier lowering is a misrepresentation of what is physically happening in the device. With the inclusion of barrier lowering, the WKB model predicts much higher tunneling current. This is where the WKB model breaks down, and may also explain why tunneling current was thought to dominate in these devices. While there are indeed some instances where tunneling current can dominate the total current flow in SFETs (Fig. 4.8), in the design cases of interest (small ϕ_B), such is not the case.

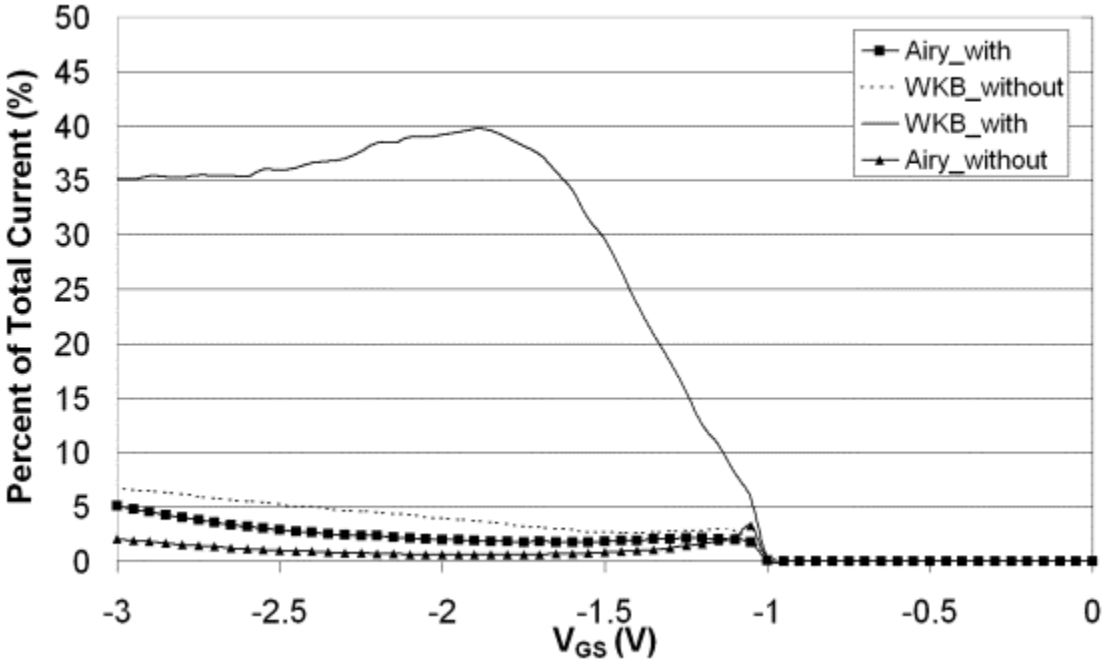

Fig. 4.7. Hole tunneling percent of total hole on state current versus gate bias for the Airy function and WKB models in the 25 nm p-channel SFET, with and without SBL. Although the non-SBL case for the WKB model is close to the Airy function model with SBL, including SBL in the WKB model gives drastically different results.

Fig. 4.7 illustrates the percent contribution of the modeled tunneling current to the total current in the 25 nm p-channel SFET. In all four cases, the thermal current is the same and modeled with the effect of SBL. As Fig. 4.7 shows, the WKB model with SBL predicts an almost equal contribution (35-40 %) to the total current as the thermal current, while the Airy function model with and without SBL predicts a much lower contribution of 2-5 %. This analysis is extended further with Fig. 4.8, which illustrates the tunneling percent of total current versus ϕ_B (equilibrium value) at $|V_{GS}|$ = 3 V when SBL is included. Fig. 4.8 shows data for both p-channel and n-channel SFETs, whereby the n-channel SFET is of the same device structure, but uses a p+ poly gate and a p-type body region instead of an n+ poly gate and an n-type body region.

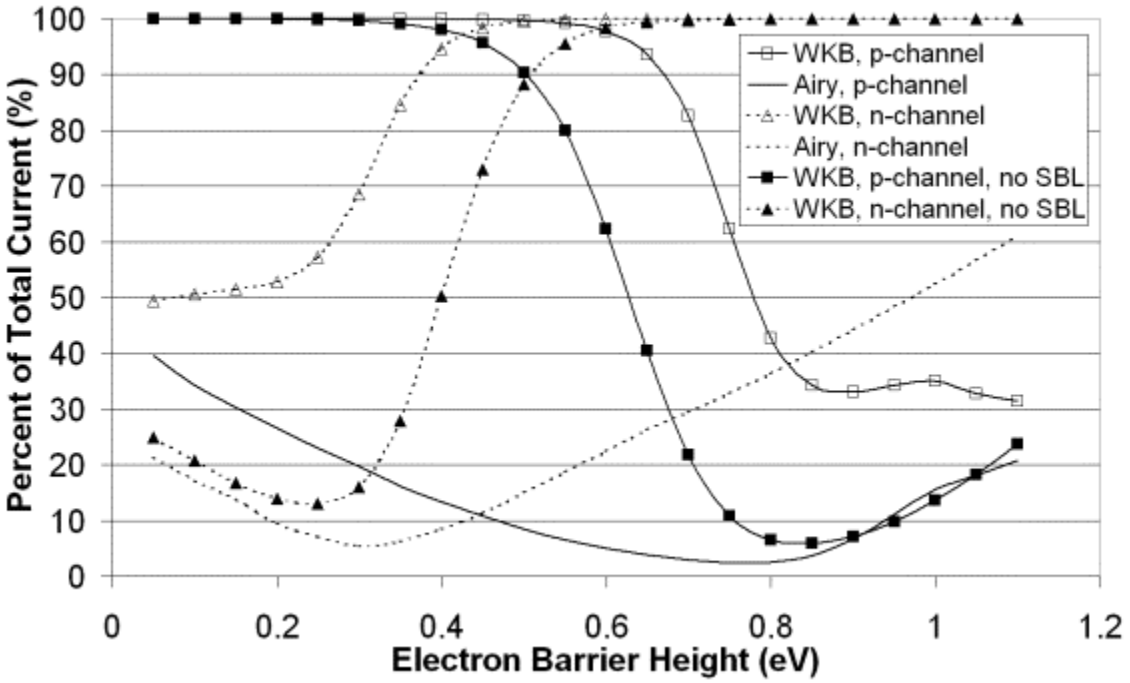

Fig. 4.8. Tunneling percent of total on-state current vs. electron SBH for 25 nm p-channel and n-channel SFETs. In the low SBH regime (~ 0.3 eV and below for electrons and ~ 0.8 eV and above for holes), the WKB model without SBL and the Airy function model with SBL predict similar contributions of tunneling current. The disparity grows considerably outside of these regions, however.

The mechanisms behind the particular form of the curves in Fig. 4.8 are not complicated. For high SBH values, the thermal current is very small, and even though SBL will decrease the SBH, the remaining Schottky barrier is still relatively large. Thus, in such an instance, Schottky barrier narrowing (which increases tunneling current) has a greater effect on increasing current than SBL. As the SBH decreases, thermal current begins to dominate, and so the tunneling contribution decreases. At low SBH values, though, the tunneling contribution becomes significant again. This is because at such low values, even though the tunneling *probability* is the same at a given energy from the top of the Schottky barrier, the tunneling *injection* increases due to the fact that the top of the Schottky barrier exists at an energy which corresponds to a larger part of the F-D distribution. In other words, while the probability of tunneling through the barrier at a given energy from the top of the barrier is the same, the probability of a carrier existing at said energy increases as the SBH decreases. Thus, more carriers are available to "try" and tunnel through the barrier, which, for a given percentage of transmission, results in more actual current.

It is interesting to note that, in Fig. 4.8 and for the particular device structure modeled, the WKB model without SBL predicts similar tunneling current to the Airy function model for low SBH values. Also, the disparity between the two models is smaller for p-channel operation than for n-channel operation. This is possibly due to the lower effective mass for electrons, for which the WKB model may be more sensitive than the Airy function model. As the SBH increases, the disparity grows tremendously, regardless of the channel type and whether or not SBL is utilized in the WKB model. For either model, however, in the low barrier height regime, it is clear that *tunneling current*

is not the dominant current mechanism. As the WKB model normally predicts a larger tunneling current injection than the Airy function model, particularly when SBL is included, one might be misled into concluding that tunneling current is the limiting factor to optimal SFET performance due to the much stronger dependence of the WKB model on SBH. While tunneling current is *a* limiting factor, though, it is not *the* limiting factor – thermal current over the Schottky barrier is the primary contributor to SFET drive current for low SBH values (Fig. 4.8), particularly in the case of the Airy function model.

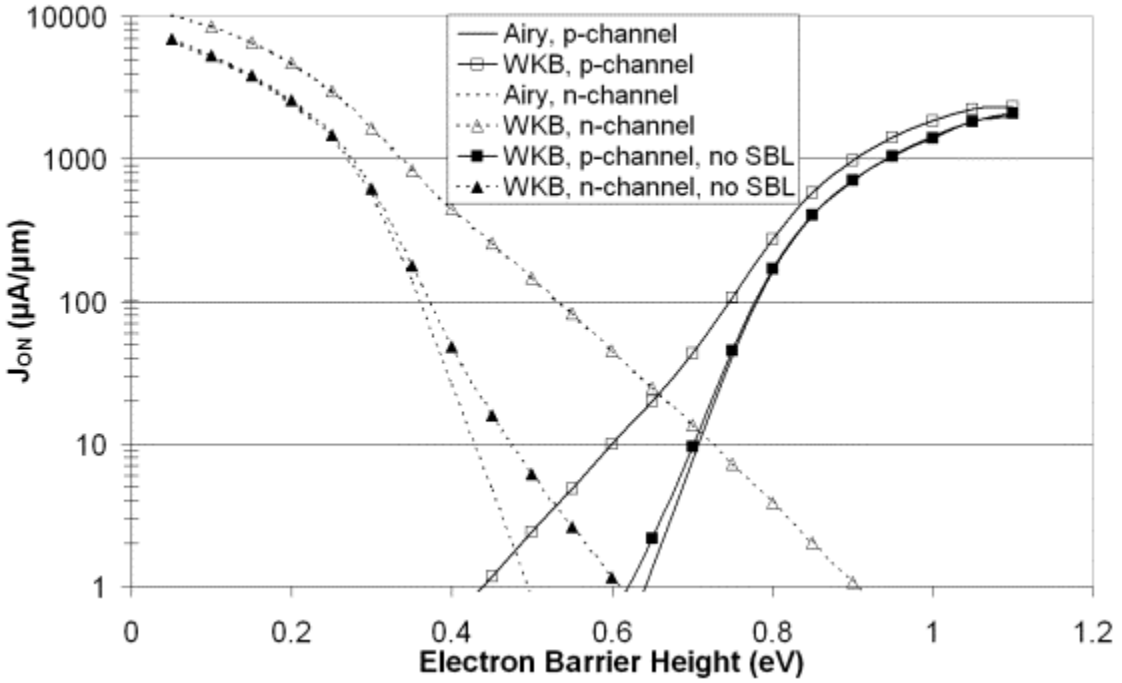

Fig. 4.9. Drive current density vs. electron SBH for the p-channel and n-channel SFETs in question, using the Airy function and WKB models. SBH was varied between 0.05 eV and 1.1 eV in 0.05 eV increments.

In observing the total modeled drive current vs. SBH for the n-channel and p-channel devices in question (Fig. 4.9), the inaccuracy of the WKB model becomes even more apparent, particularly for the n-channel device. As the SBH approaches midgap values, the WKB model with SBL predicts a much higher level of tunneling current,

which actually starts to dominate over thermal current as Fig. 4.8 also shows. For SBH values around 0.4 eV and below for electrons and around midgap and below for holes (which is midgap and above for the electron SBH), the Airy function model with SBL and the WKB model without SBL result in a surprisingly good agreement. However, while one might argue that the good agreement between the WKB model without SBL and the Airy function model with SBL warrants the use of the WKB model due to its relative mathematical simplicity over the Airy function approach, again, the exclusion of SBL is a misrepresentation of what is physically happening in the device.

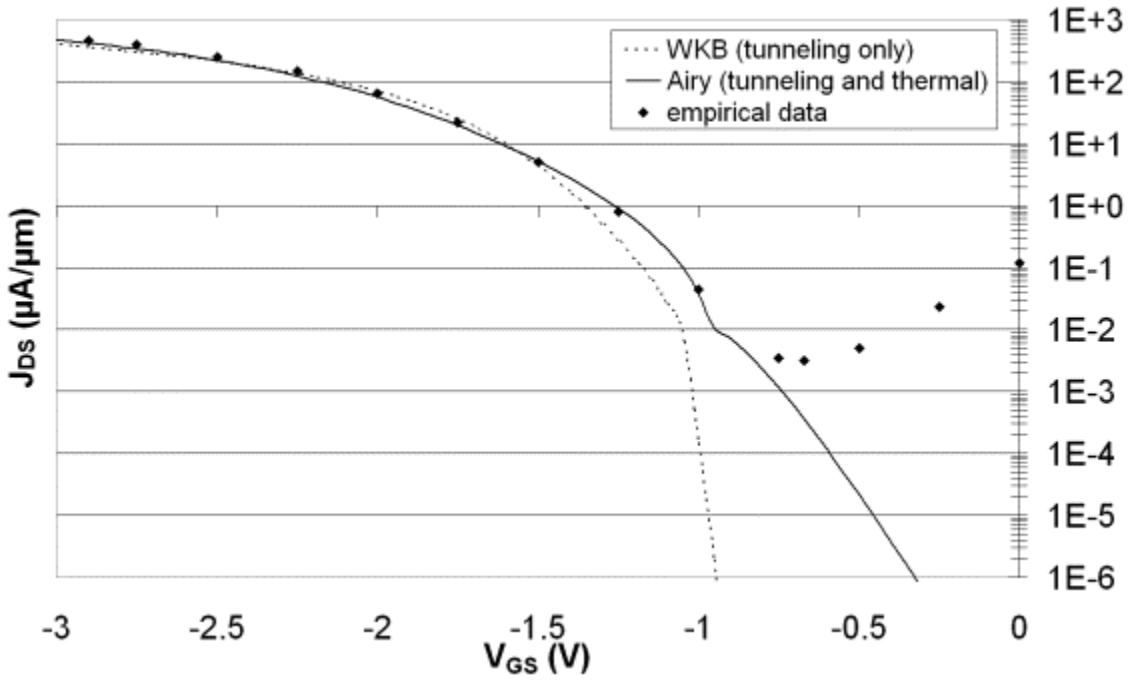

Fig. 4.10. Comparison of Airy function model with the inclusion of thermal current and a particular utilization of the WKB model without thermal current to empirical data for the 25 nm p-channel SFET modeled throughout this section.

If the WKB model is utilized with SBL, without "shifting" the previously discussed integration limits (the overestimated case discussed previously), and without accounting for thermal current over the Schottky barrier, a surprisingly close fit to empirical data, as well as to the Airy function model with the inclusion of thermal

current, can be achieved at and near the on state. This is shown in Fig. 4.10, in which the p-channel SFET is modeled, and the empirical data extracted from [16]. The increase in current at low gate biases in the empirical data is due to electron tunneling injection at the drain. As only inversion carrier current is of interest in this section, neither model used here fully fits the data. However, a very good fit with the Airy function model has been achieved over the entire V_{GS} range, and is covered in Section 4.6. In any case, it is important to note that caution must be taken in choosing a particular tunneling model, as the fundamental implications of the results can be misleading. While the WKB model is useful for relatively large tunnel barriers such as the source-to-drain barrier during contact potential modulation (low V_{GS} range in Figs. 4.1 and 4.6), its utility is diminished in the application to Schottky barrier modulation, where the lateral electric field induced by the gate is relatively large and the consequent tunnel barrier is very narrow.

4.8 Device Optimization: Conventional SFETs

Optimizing thermal current over the Schottky barrier, interestingly enough, also happens to optimize tunneling current through the barrier, so again one must take caution in interpreting their observations of increased or decreased current in a particular device structure. Ultimately, what is required for optimal SFET performance is a small Schottky barrier, and not just in terms of height – the barrier width must also be small. A smaller barrier width implies greater SBL due to the higher lateral field, which decreases the barrier height and increases thermal current. This also has the added benefit of increasing tunneling injection. In terms of integrating this requirement into an actual device, it has been suggested that the electrostatic effects in ultrathin body SOI substrates increase the

electric field at the Schottky source/drain regions, as the gate assumes greater control over the potential within the entire body region [6].

Another approach to increasing SFET performance is also possible, whereby ultrashallow halo regions of very high dopant concentration and lateral abruptness can be formed to induce a large field at the M-S junctions [17]. While these two approaches modify the barrier width, it has also been shown that the actual height of the Schottky barrier can be "tuned" to very low values for electrons by implanting valence-mending adsorbates before metal deposition and subsequent silicide formation [18]. Whether such an approach can or will be demonstrated for holes, however, remains to be seen. For the remainder of this section, unless otherwise noted, model results utilize the Airy function tunneling model in the overestimated case discussed previously.

For a given device structure, modulating the SBH modulates the accumulation carrier injection at the drain as well as the inversion channel drive current, as Fig. 4.11 shows for the 25 nm p-channel structure modeled throughout this chapter. Also, as the electron barrier height (ϕ_{Bn}) decreases, the trough of the transfer curve is shifted in the negative V_{GS} direction. In the case of the results of Fig. 4.11, switching gate from an n+ poly gate to a p+ poly gate would place this trough at V_{GS} = 0 V for ϕ_{Bn} values around 0.65 eV. Conveniently, NiSi exhibits this value [18]. Likewise for the PtSi source/drain device, using a fully-silicided (FUSI) gate would shift the trough of the transfer characteristic to about 0 V. Both of these curves are shown in Fig. 4.12, where electron tunneling injection at the drain is significantly reduced at V_{GS} = 0 V (poly depletion is not included in this model – only the effect of the gate workfunction shift is accounted for). The NiSi implementation exhibits a drive current of 31 μA/μm, which, while respectable,

is insufficient for such small devices. The FUSI PtSi implementation not only responds better to V_{GS}, but also exhibits a drive current of 1.09 mA/μm – a considerable gain over the original implementation (498 μA/μm). This clearly places a performance preference toward PtSi over the much less expensive NiSi for PFET operation.

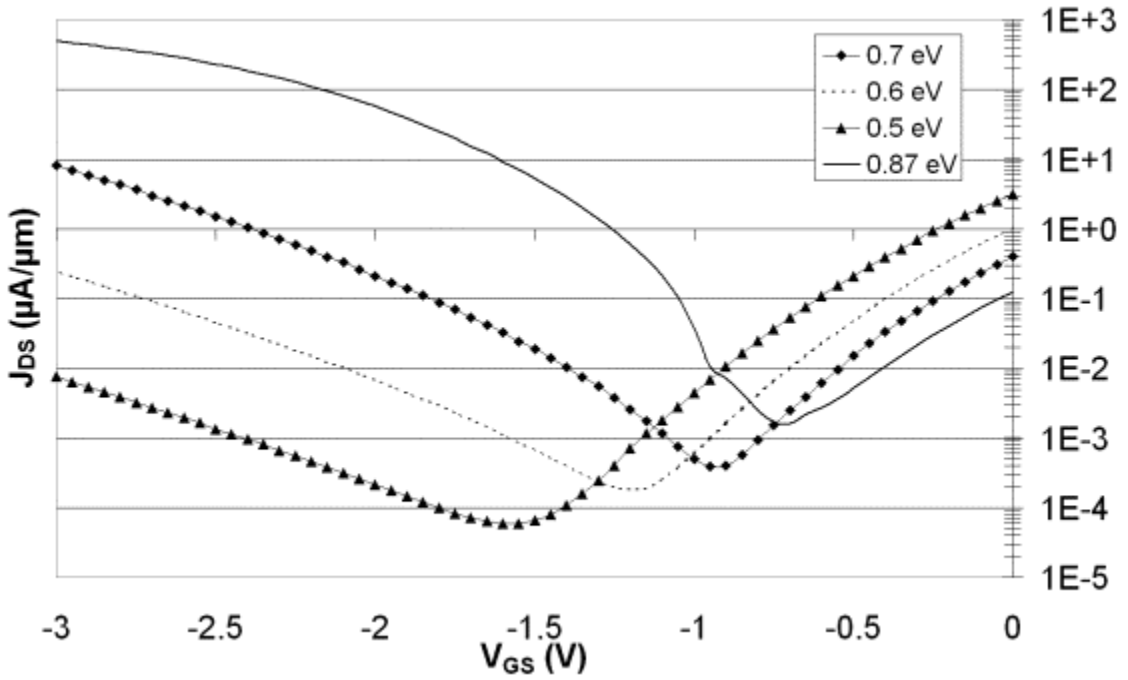

Fig. 4.11. Modeled transfer characteristics for modified versions of the 25 nm p-channel SFET from Fig. 4.2. Subsurface leakage was excluded, and barrier heights in the legend are to electrons. The 0.87 eV curve is the model result from Fig. 4.2.

It is also interesting to note that, for the same device design in Fig. 4.11, changing the SBH has little if any effect on the shape of the transfer characteristic, for it is mostly just shifted in one direction or another. The subthresold swing also changes very little if at all, and intuitively this would make sense since SBL is largely independent of SBH (there is actually a small dependence, as the field at the M-S interface is partly a function of SBH, but this dependence is not very large). Considering how a change in thermal current is only dependent on SBL for a change in V_{GS} (assuming ballistic transport), if a

device structure does demonstrate a significant change in subthreshold swing with SBH, then this change can be attributed to tunneling current; again, as (10) and (13)-(18) show, tunneling current has a dependence on SBH *and* the barrier width at a given energy.

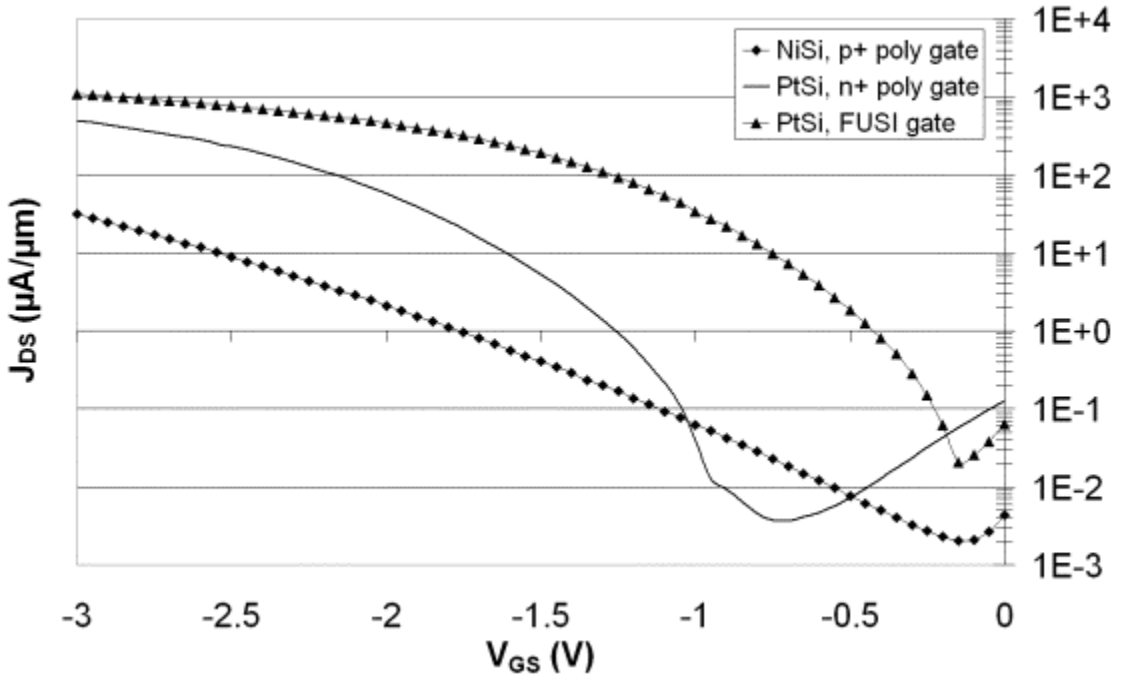

Fig. 4.12. Modeled transfer characteristics for modified versions of the 25 nm p-channel SFET from [16]. Subsurface leakage was set to 1 nA/µm. Electron injection at the drain can be reduced to values less than or equal to the subsurface leakage when the appropriate gate workfunctions are utilized.

It turns out that NiSi still has potential in SFETs for both "conventional" and bulk switching designs, despite the modeling result in Fig. 4.12. From a purely SBH perspective, it was shown in [18] that the SBH from NiSi to electrons can be reduced to values as low as 0.07 eV via implantation of moderate doses of S^+ before the Ni is deposited and the silicide is formed. Properly tuned NiSi can therefore serve as an excellent material for *n-channel* SFETs, thus replacing $ErSi_2$ [13], [15] ($\phi_{Bn} \sim 0.25$ eV) as the most promising material of choice. As Ni is a more common material than Er, a shift to NiSi over $ErSi_2$ can also reduce fabrication costs. Transfer curves for a 25 nm $ErSi_2$

device and a tuned NiSi device (same basic structure as the device from Fig. 4.2) are shown in Fig. 4.13. For the n-channel device structure, α and η are 1.1 and 2.6 for electrons (inversion carriers in this case), respectively, and 0.1 and 1.1 for holes (accumulation carriers in this case), respectively – the exact opposite of the values used for the p-channel device modeled throughout this chapter (these values were also used previously in this chapter for n-channel investigations). Drive currents for the $ErSi_2$ devices with the p+ poly gate and FUSI gate, as well as the tuned NiSi device, are modeled as 1.46 mA/μm, 3.31 mA/μm, and 8.91 mA/μm, respectively.

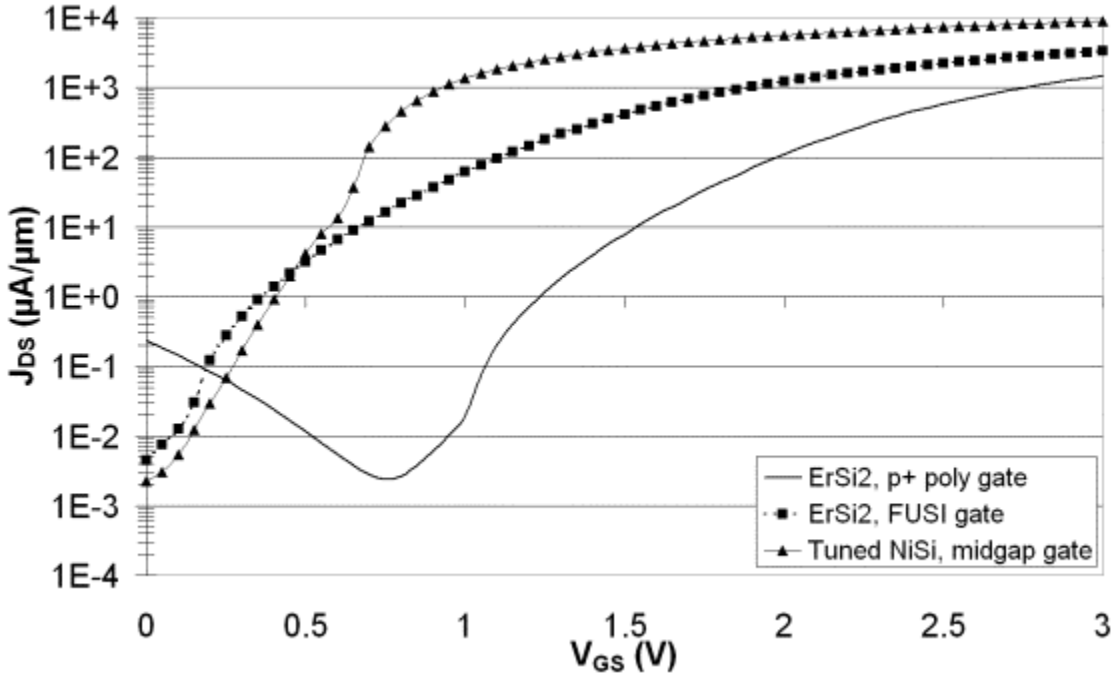

Fig. 4.13. Modeled transfer characteristics for 25 nm n-channel SFETs using $ErSi_2$ and tuned NiSi source/drain regions. Subsurface leakage was set to 1 nA/μm. The tuned NiSi device with the midgap gate exhibits superior performance to the $ErSi_2$ device.

For the tuned NiSi device with the midgap gate in Fig. 4.13, a lower leakage and higher drive current are achieved over the $ErSi_2$ device with the FUSI gate. Also, the tuned NiSi device exhibits a low subthreshold swing over a larger V_{GS} range, resulting in

a larger V_{sbfb} (Fig. 4.14), lower operating voltage capability, and higher SCE immunity. It is also interesting to note that SBL results in a *negative* SBH as the gate-induced electric field increases, which meets the recommendation of [7] for high performance SFETs. Even though tuned NiSi exhibits a lower SBH than $ErSi_2$ (0.07 eV vs. 0.25 eV), lower leakage is possible via gate workfunction engineering; however, one would expect the subsurface leakage to be higher in the tuned NiSi device, and so its advantage over $ErSi_2$ may only be realized with ultrathin body devices (i.e., FinFETs, UTBSOI, etc.).

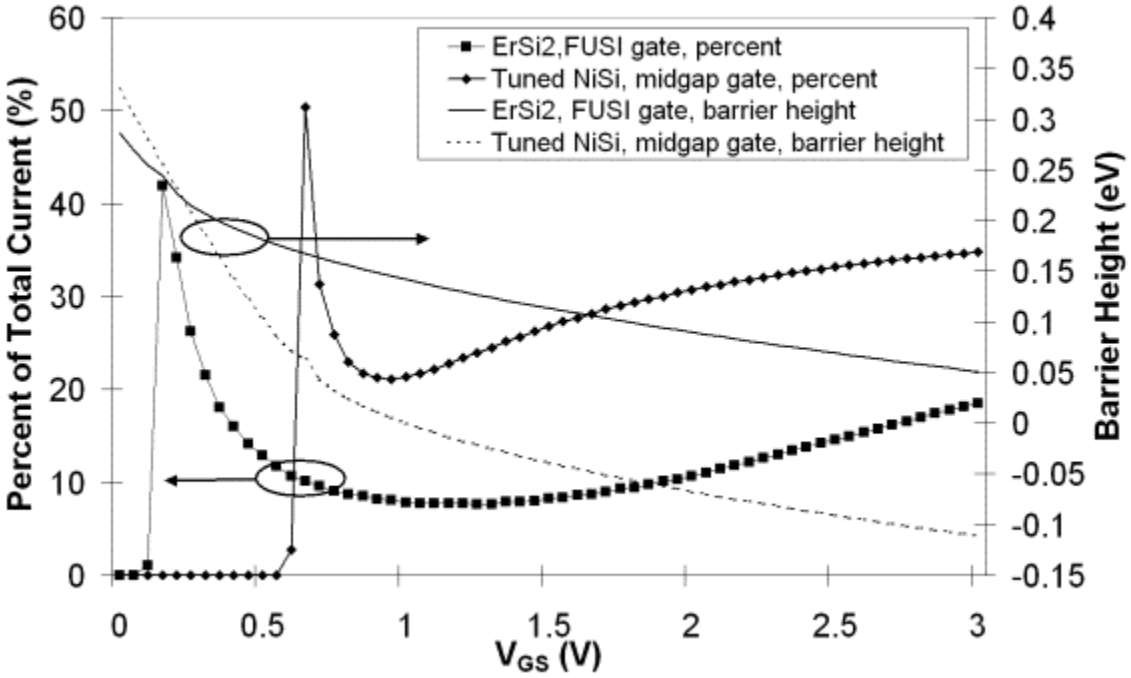

Fig. 4.14. Model results for electron barrier height and electron tunneling percent of total current vs. V_{GS} for 25 nm p-channel SFETs with $ErSi_2$ and tuned NiSi source/drain regions. Subsurface leakage was set to 1 nA/μm.

4.9 Device Optimization: Bulk Switching SFETs

From the perspective of bulk switching SFETs, NiSi is a promising material of choice because the bulk switching design allows it to be used for both p-channel and n-channel devices. As the primary feature of the bulk switching device is the halo region

next to the source/drain silicide, the lateral electric field at the M-S junctions determines the maximum current injection that a particular silicide can support. Under equilibrium, this lateral field is defined by the halo dopant concentration, and so for a given V_{DS}, one can determine the combinations of SBH and halo dopant concentration that result in the desired maximum level of current injection that the source/drain silicide can support. To solve for this requires solving for Poisson's Equation in the semiconductor. First, some assumptions are made in accordance to the conditions of a bulk switching SFET that allow for a simpler solution. It is assumed that the halo regions are degenerately doped and by virtue of this: 1.) the surface potential, Ψ_s, at the M-S interface is approximately equal to the SBH ($E_F \sim E_c$ far from the M-S interface); 2.) the minority carrier concentration is negligible. With this in mind, the charge density in an n-type semiconductor as a function of distance from the M-S junction, $\rho(x)$, can be expressed as:

$$\rho(x) = q(N_d - n(x)) \tag{23}$$

where $n(x)$ is the electron concentration as a function of distance from the M-S junction, which can be expressed as:

$$n(x) = n_0 \exp\left(\frac{q\Psi(x)}{kT}\right) \tag{24}$$

where n_0 is the equilibrium electron concentration (which can be treated as N_d at 300 K) and $\Psi(x)$ is the potential with respect to the "bulk" as a function of distance from the M-S junction. In this analysis, $\Psi(x)$ is treated as increasing when it approaches the conduction band and decreasing when it approaches the valence band. Plugging (24) into (23) gives (25).

$$\rho(y) = q\left(N_d - N_d \exp\left(q\frac{\Psi(x)}{kT} \right) \right) \tag{25}$$

Poisson's Equation is expressed in (26), and when combined with (25), gives the expression in (27), where ε_s is the relative permittivity of the semiconductor.

$$\frac{d^2\Psi(x)}{dx^2} = \frac{-\rho}{\varepsilon_0 \varepsilon_s} \tag{26}$$

$$\frac{d^2\Psi(x)}{dx^2} = \frac{-qN_d}{\varepsilon_0 \varepsilon_s}\left(1 - \exp\left(\frac{q\Psi(x)}{kT} \right) \right) \tag{27}$$

At this point, some mathematical savvy is utilized to provide a relatively simple solution to the electric field at the M-S junction. If both sides of (27) are multiplied by $2*d\Psi(x)/dx$, the left side of (27) becomes (28), which, when expanded, gives (29). Working backwards from the Product Rule of differentiation, (29) simplifies to (30). Now, (27) can be expressed as (31).

$$\frac{2d\Psi(x)}{dx}\frac{d^2\Psi(x)}{dx^2} \tag{28}$$

$$\frac{d\Psi(x)}{dx}\frac{d^2\Psi(x)}{dx^2} + \frac{d\Psi(x)}{dx}\frac{d^2\Psi(x)}{dx^2} \tag{29}$$

$$\frac{d}{dx}\left(\frac{d\Psi(x)}{dx} \right)^2 \tag{30}$$

$$\frac{d}{dx}\left(\frac{d\Psi(x)}{dx} \right)^2 = \frac{-2qN_d}{\varepsilon_0 \varepsilon_s}\left(1 - \exp\left(\frac{q\Psi(x)}{kT} \right) \right)\frac{d\Psi(x)}{dx} \tag{31}$$

By converting $\Psi(x)$ to Ψ and integrating both sides of (31) with respect to x over the range where $\Psi = \Psi_s$ ($x = 0$) to where $\Psi = 0$ far from the M-S junction gives the expression in (32). Since $d\Psi/dx = 0$ when $\Psi = 0$, the integral in (32) results in the

expression in (33). The square root of the expression in (33) gives the electric field at the surface, ξ_s. Recalling one of the initial assumptions, whereby degenerate doping results in $q\Psi_s \sim \phi_B$, the solution to (33) can be expressed as (34). An equivalent p-type semiconductor approach gives the same result when the $\Psi(x)$ convention is reversed.

$$\left(\frac{d\Psi}{dx}\right)^2 = \frac{2qN_d}{\varepsilon_0\varepsilon_s}\int_{\Psi_s}^{0}\left(1-\exp\left(\frac{q\Psi}{kT}\right)\right)d\Psi \tag{32}$$

$$\left(\frac{d\Psi}{dx}\right)^2 = \frac{2qN_d}{\varepsilon_0\varepsilon_{si}}\left(\frac{kT}{q}\left(1-\exp\left(\frac{q\Psi}{kT}\right)\right)+q\Psi\right)\Big|_{\Psi_s}^{0} \tag{33}$$

$$\xi_{s,n} = \sqrt{\frac{2qN_d}{\varepsilon_0\varepsilon_{si}}\left(\frac{kT}{q}\left(\exp\left(\frac{q\phi_B}{kT}\right)-1\right)-\phi_B\right)} \tag{34}$$

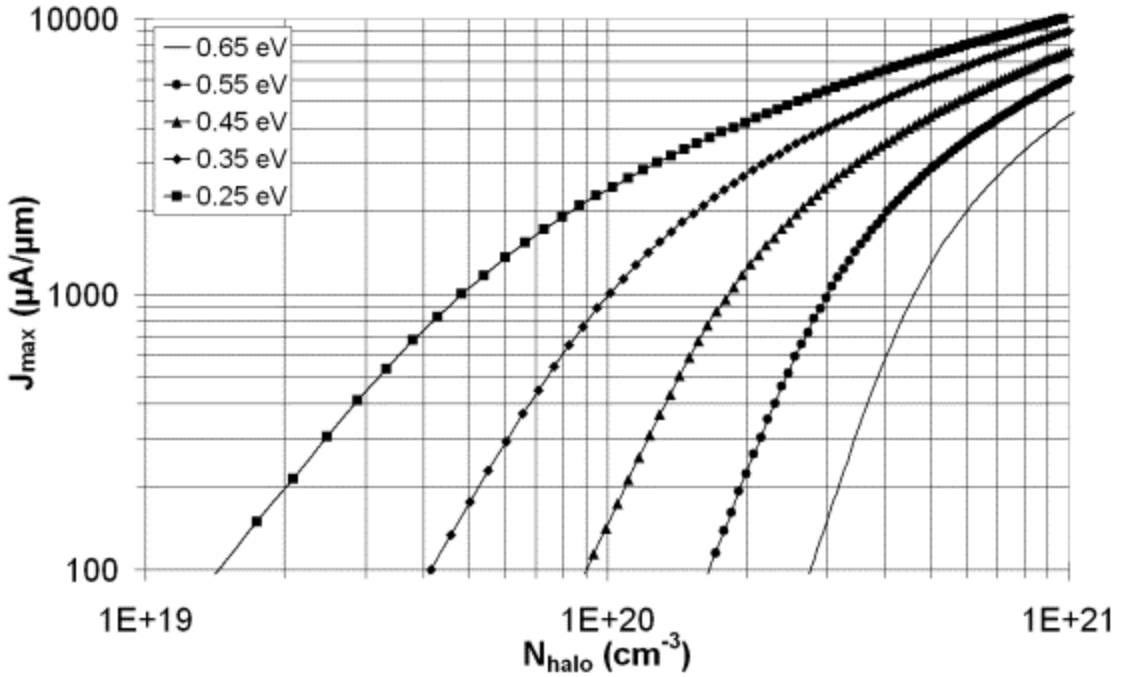

Fig. 4.15. Maximum current density through a Schottky barrier to electrons versus halo dopant concentration for various SBH values at V_{DS} = 1.1 V.

Figs. 4.15 and 4.16 show the maximum predicted current density through Schottky barriers to electrons and holes, respectively, as a function of halo dopant concentration for $|V_{DS}|$ = 1.1 V. In the case of NiSi, which presents an SBH to electrons of 0.65 eV (and therefore a hole SBH of ~ 0.47 eV), halo dopant concentrations in the high $1x10^{20}$ cm^{-3} range can support current densities of over 2 mA/μm for both n-channel and p-channel devices at 1.1 V.

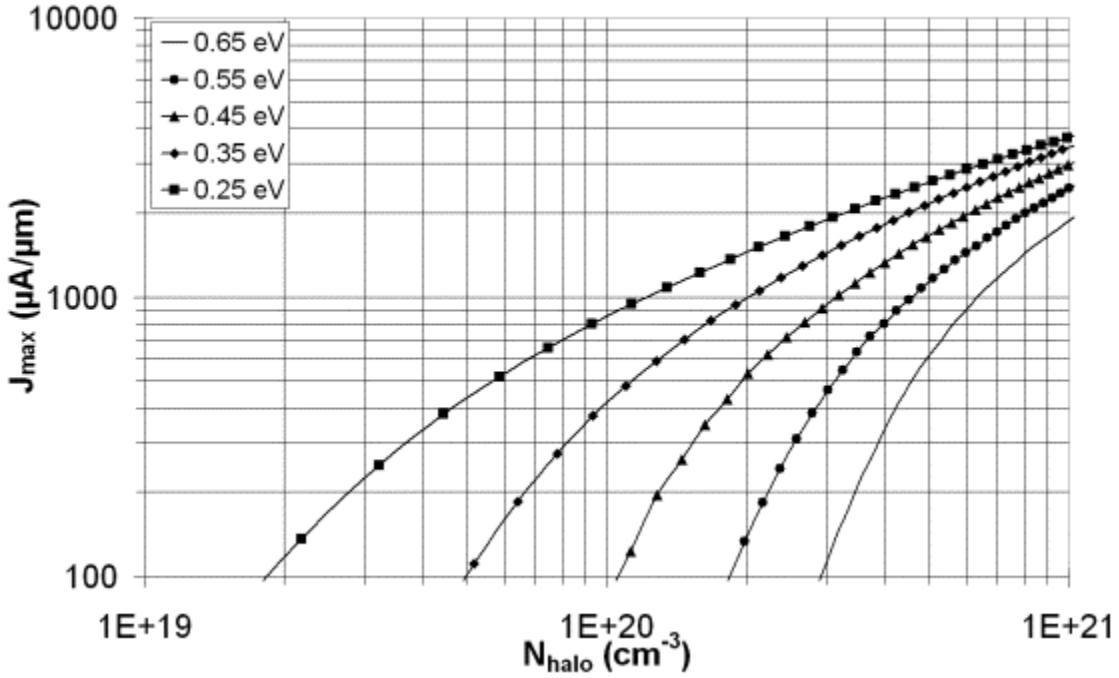

Fig. 4.16. Maximum current density through a Schottky barrier to holes versus halo dopant concentration for various SBH values at V_{DS} = -1.1 V.

The 2005 ITRS Process Integration, Devices & Structures Roadmap projects V_{DD} to be 1.1V for 25 nm physical gate lengths (as was modeled throughout this chapter) and 1.51 mA/μm NMOS drive current density for high performance logic [19]. At a first glance, then, it would seem that even though NiSi provides a relatively large SBH to both electrons and holes in its non-tuned form, drive current would not be limited by the SBH under the condition of high concentration halo regions (on the order of the source/drain

dopant level for conventional MOSFETs). As mentioned in Chapter 3, these halo regions can be formed by an implant through silicide (ITS) process, which may indeed facilitate such current levels. This is currently speculative, however, and for two reasons. First, the assumptions made to result in Figs. 4.15 and 4.16 are approximations (especially the first assumption, which is effectively a constant V_{bi}, as well as the fact that bandgap narrowing was excluded and the definition of α is somewhat arbitrary), and so the results shown are likely an underestimation of the expected current levels in degenerately doped semiconductors. Second, the post-ITS anneal for NiSi is performed at around 600 °C [17], for which the solid solubility of boron and phosphorus in silicon are *not* in the high-$1x10^{20}$ range if one extrapolates from existing data [20]. However, this is a trickier notion, as low temperature dopant activation of boron and phosphorus is not well understood, especially the role that a silicide diffusion source may play in such activation. Any potential insufficiency in current injection due to solid solubility limits or some other limitation can be mitigated by adjusting the gate influence on the M-S junctions (i.e., the gate overlap), although this imposes the ubiquitous tradeoff between drive current and off state leakage, as well as overlap capacitance.

Beyond the Schottky barrier, current may be degraded or limited by the reduced mobility within a very thin SOI body region or by velocity saturation near the drain. Mobility degradation via the halo regions should not noticeably affect drive current, as these devices effectively operate in accumulation mode. Thus, the mathematical contribution of the halo dopant concentration to decreasing the resistivity (and hence resistance) of the halo region far outweighs the induced mobility reduction in this region. Certainly, such a reduction in channel resistance would be compounded with high

mobility substrates such as strained silicon or $Si_{1-x}Ge_x$.

Optimizing the use of halo regions in bulk switching SFETs requires, in addition to high dopant concentrations, highly abrupt junctions and an optimal size of the halo region as a percentage of the channel region. If the halo region is too large, although drive current is increased, the device is more susceptible to DIBL at the source due to the smaller lightly doped/undoped body region for a given gate overlap/underlap to the M-S junction (i.e., the halo constitutes a higher percentage of the body region). If the halo region formed by ITS is too small, drive current is reduced, and this reduction depends on the abruptness of the halo region which, for small halo regions, determines the dopant concentration at the M-S interface.

Determining an appropriate size for the halo regions (and thus the post-ITS anneal thermal cycle) requires, at the very least, some insight into the depletion region extending from the M-S junction into the halo region at the maximum operating voltage. Starting with Pierret's derivation of the depletion width at an M-S junction [12], and again using the assumption that for degenerately doped semiconductors, $V_{bi} \sim \phi_B$, one reaches the expression in (35), where N_{halo} is the halo dopant concentration (assuming a halo region of uniform doping).

$$W_D = \sqrt{\frac{2\varepsilon_0\varepsilon_s}{qN_{halo}}\left(\phi_B + V_{DD}\right)} \tag{35}$$

Fig. 4.17 shows W_D versus N_{halo} at various V_{DD} and SBH values. The SBH was varied from 0.25 eV (smallest W_D) to 0.65 eV (largest W_D) in 0.1 eV increments for each V_{DD} value. As one might expect, W_D has a small dependence on SBH and V_{DD} if N_{halo} is

large enough (high $1x10^{20}$ cm^{-3} range). Going back to Figs. 4.15 and 4.16, for maximum current densities on the order of 2 mA/μm for both the n-channel and p-channel devices, N_{halo} ~ $6\text{-}8x10^{20}$ cm^{-3}, which corresponds to W_D between ~ 0.85 nm and 1 nm at V_{DD} = 1.1V. This implies that, for a 25 nm device, the halo region must be at least 1 nm wide with the aforementioned concentration range to prevent punchthrough within the halo region. Assuming that the halo region is 2.5 nm wide at each end of the device, and that the M-S junctions are very close to the gate edge (i.e., the gate overlap/underlap is approximately zero), the source and drain halo regions combined constitute 20 % of the channel region, which means that 80 % of the channel region can be lightly doped (or undoped), which considerably improves channel mobility in comparison to modern conventional MOSFETs.

Such a condition, however, also assumes that quantum carrier confinement within the halo region does not take place. With a very narrow halo region (on the order of a few nm), a quantum well is formed between the Schottky barrier and the halo-body thermal barrier. Even for high halo dopant concentrations, the entire halo region becomes depleted under these conditions due to the reduction in the majority carrier concentration. This should not affect majority carrier thermal leakage over the halo-body thermal barrier; however, minority carrier thermal leakage over and tunneling leakage through said barrier should increase, thus implying that halo depletion occurs at a halo width larger than what conventional theory, such as (35), would predict. This in turn implies an optimal halo width that is large enough to minimize minority carrier leakage but also small enough to optimize SCE immunity for a given source-to-drain spacing.

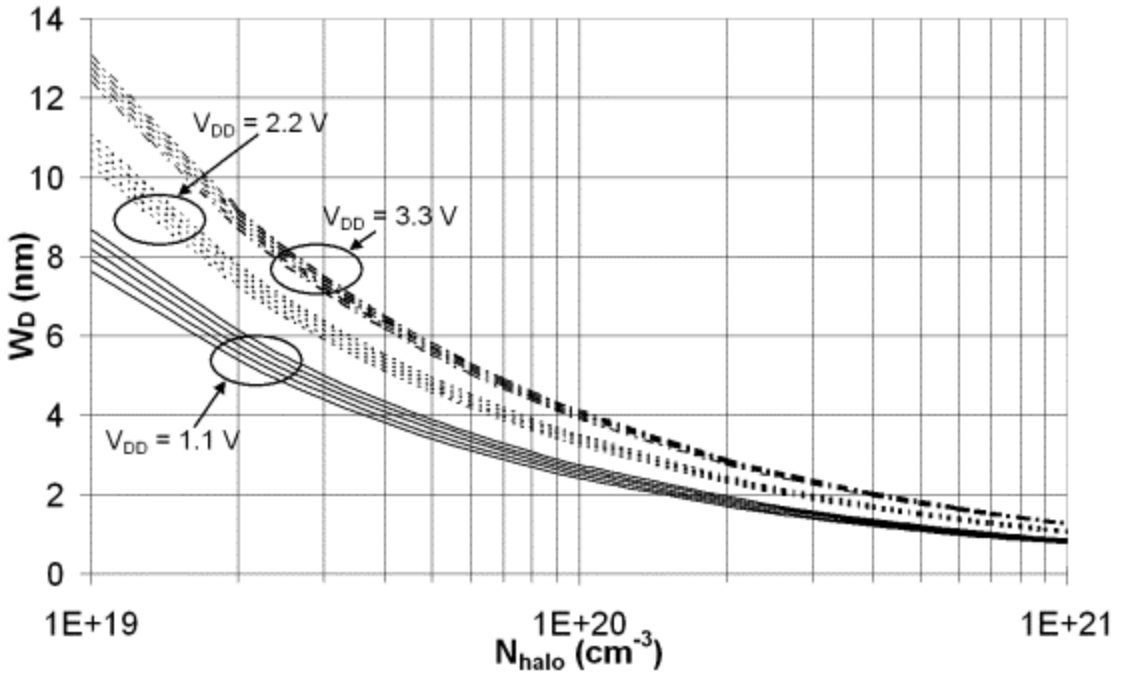

Fig. 4.17. Depletion width at an M-S junction versus N_{halo}, SBH, and V_{DD}. The SBH was varied from 0.25 eV (smallest W_D) to 0.65 eV (largest W_D) in 0.1 eV increments for each V_{DD} value.

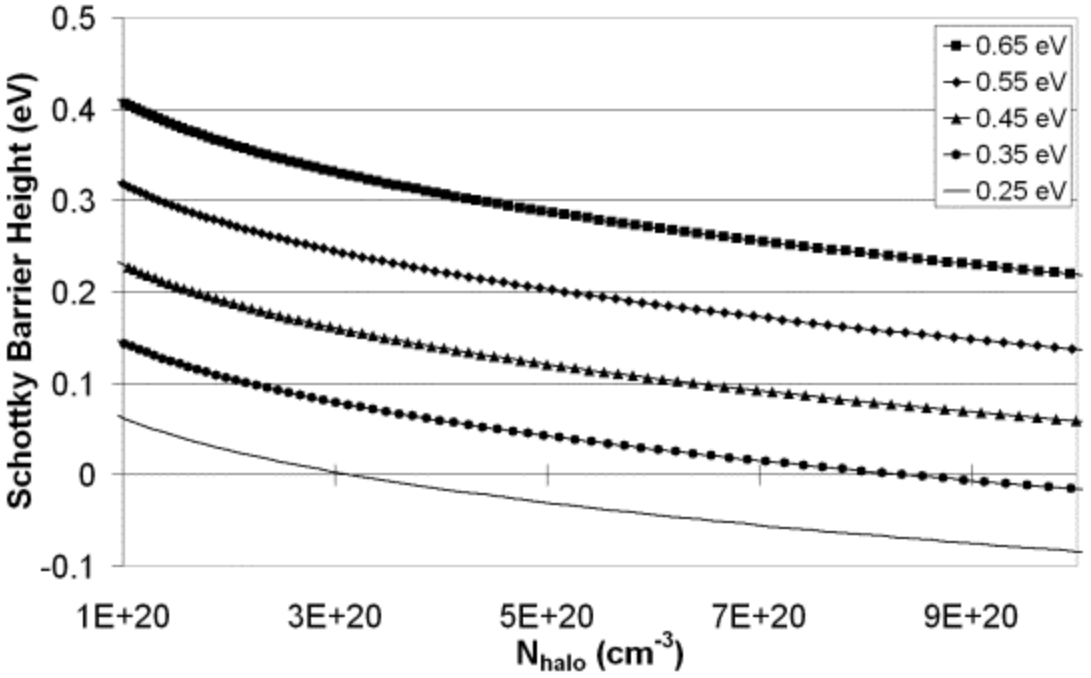

Fig. 4.18. SBH versus N_{halo} for various equilibrium SBH values when SBL is included. In this case, α is set to 1.1 so as to be consistent with the device structure modeled throughout this chapter.

As was mentioned in Chapter 3, the lateral field induced by the high N_{halo} results in SBL, which accounts for a considerable portion of the increase in maximum current injection through and over the Schottky barrier as N_{halo} is increased. The resultant SBH values for given equilibrium SBH values are shown in Fig. 4.18 for N_{halo} between $1x10^{20}$ and $1x10^{21}$ cm^{-3}. For NiSi to n-type silicon, for example, the equilibrium value of 0.65 eV is actually lowered to about 0.25 eV in the presumed N_{halo} range of interest for 25 nm bulk switching SFETs (~ $6\text{-}8x10^{20}$ cm^{-3}). If the equilibrium SBH is low enough, the resultant SBH can actually turn negative, as Fig. 4.18 also shows.

It is noted that, in Fig. 4.18, α was set to 1.1 rather than the "classical" value of 1, so as to be consistent with the 25 nm devices modeled earlier in this chapter. Interestingly, it turns out that the higher than expected SBL is not an unreasonable proposition. Shenai and Dutton have shown that SBL has a contribution from Heine tail decay (dipole lowering) as well as from the image force [13], and that SBL is thus considerably larger than what is predicted classically. While the model approach for SBL in [13] was more physically rigorous than the approach taken here, Andrews and Lepselter achieved similar results to [13] by adding an empirical fitting parameter (also called α, but used differently) to the conventional SBL equation [14], shown in (36), the form of which is fairly similar to that of (19). These results are shown in Fig. 4.19. Therefore, while the physical meaning of α in (19) is somewhat ambiguous in nature, it can be considered at least in part to account for the effect of Heine tail decay on a very superficial level.

$$\Delta\phi_B = \sqrt{\frac{q|\xi|}{4\pi\varepsilon_0\varepsilon_{si}}} + \alpha|\xi| \tag{36}$$

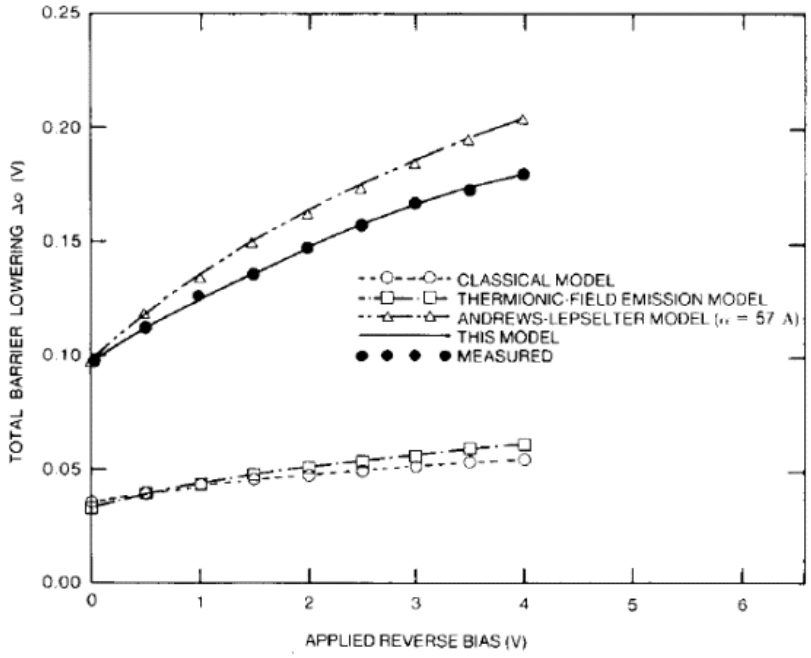

Fig. 4.19. Comparison of Andrews-Lepselter model [14], Shenai-Dutton model [13], classical model, and thermionic field emission model for SBL for an Al-nGaAs Schottky diode in reverse bias, adapted from [13]. "This model" refers to the Shenai-Dutton model. A considerable difference exists between conventional theory and the models from [13] and [14].

Reverting back to bulk switching SFET design, optimizing the off state characteristics in bulk switching SFETs is perhaps equally challenging as optimizing the on state characteristics. The off state thermal barrier is that which exists between the halo region and the lightly doped/undoped body region. One might suggest using an opposite dopant type of moderate to high concentration for the body region than the halo region to increase the off state barrier. However, while the off state thermal barrier is larger, so are the source-body and drain-body *junction capacitances*. This reduces the high frequency performance of the device due to increased coupling at the source-body and drain-body junctions. Using an undoped or lightly doped body, in comparison, would decrease the

junction capacitances by several orders of magnitude. Also as a result of using an opposite dopant for the body region, the built-in field at the drain becomes larger, which results in GIDL at a lower V_{DS}. Ultimately, the device becomes a conventional partially depleted SOI (PD SOI) MOSFET with fully silicided source/drain regions, which results in additional design considerations such as the history effect and the kink effect.

A more effective method of increasing the off state thermal barrier would be to use a lightly doped or undoped body region and to *modify the gate workfunction*. The immediate advantage here, particularly for the undoped body, is a reduction in discrete dopant effects due to statistical variations in dopant concentration, thus decreasing the overall statistical variation in device performance across a given chip. While this requires thin body regions such that a change in the gate workfunction affects the *total* leakage within the device (i.e., there is little or no subsurface leakage), the effect of greater gate control over current through the device is achieved. Such an implementation might best be referred to as "pseudo-FD SOI," as while the lightly-doped/undoped body region is fully depleted, the halo regions should remain largely intact. For very small devices (i.e., 25 nm and below), metallic gates are attractive due to the reduction in *EOT* as poly depletion is eliminated. For the simplest process, one might consider using the same NiSi for the n-channel and p-channel devices as the gate material for both devices (FUSI gates), as it is roughly a midgap material (but with a slight bias toward the p-type region of the semiconductor bandgap). In such a case, the full off state of both devices might not be reached, although it would be possible to individually tune the gate workfunctions to achieve the desired result using methods such as silicidation-induced impurity segregation (SIIS) [21], [22].

Chapter 4 References

[1] Z.H. Liu, C. Hu, J.H. Huang, T.Y. Chan, M.C. Jeng, P.K. Ko, Y.C. Cheng, "Threshold Voltage Model for Deep-Submicrometer MOSFET's," *IEEE Trans. Elec. Dev.*, Vol. 40, no. 1, 1993, pp. 86-95.

[2] B. Winstead, U. Ravaioli, "Simulation of Schottky Barrier MOSFETs with a Coupled Quantum Injection/Monte Carlo Technique," *IEEE Trans. Elec. Dev.*, 2000, Vol. 47, no. 6, pp. 1241-1246.

[3] R.A. Vega, "On the Modeling and Design of Schottky Field Effect Transistors," *IEEE Trans. Elec. Dev.*, Vol. 53, no. 4, 2006, pp. 866-874.

[4] S. Xiong, T.J. King, J. Bokor, "A Comparison Study of Symmetric Ultrathin-Body Double-Gate Devices With Metal Source/Drain and Doped Source Drain," *IEEE Trans. Elec. Dev.*, Vol. 52, no. 8, 2005, pp. 1859-1867.

[5] K. Matsuzawa, K. Uchida, A. Nishiyama, "A Unified Simulation of Schottky and Ohmic Contacts," *IEEE Trans. Elec. Dev.*, Vol. 47, no. 1, 2000, pp. 103-108.

[6] J. Knoch, J. Appenzeller, "Impact of the channel thickness on the performance of Schottky barrier metal-oxide-semiconductor field-effect transistors," *App. Phys. Lett.*, Vol .81, no. 16, 2002, pp. 3082-3084.

[7] J. Guo, M.S. Lundstrom, "A Computational Study of Thin-Body, Double-Gate, Schottky Barrier MOSFETs," *IEEE Trans. Elec. Dev.*, Vol. 49, no. 11, 2002, pp. 1897-1902.

[8] M. Bescond, J.L. Autran, D. Munteanu, N. Cavassilas, M. Lannoo, "Atomic-scale Modeling of Source-to-Drain Tunneling in Ultimate Schottky Barrier Double-Gate MOSFET's," *33rd Conference on European Solid-State Device Research*, 2003, pp. 395-398.

[9] K.F. Brennan, C.J. Summers, "Theory of resonant tunneling in a variably spaced multiquantum well structure: An Airy function approach," *J. Appl. Phys.*, Vol. 61, no. 2, 1987, pp. 614-623.

[10] S. Zhu, H.Y. Yu, S.J. Whang, J.H. Chen, C. Shen, C. Zhu, S.J. Lee, M.F. Li, D.S.H. Chan, W.J. Yoo, A. Du, C.H. Tung, J. Singh, A. Chin, D.L. Kwong, "Schottky-Barrier S/D MOSFETs With High-K Gate Dielectrics and Metal-Gate Electrode," *IEEE Elec. Dev. Lett.*, Vol. 25, no. 5, 2004, pp. 268-270.

[11] R.A. Vega, "A Comparison Study of Tunneling Models for Schottky Field Effect Transistors and the Effect of Schottky Barrier Lowering," *IEEE Trans. Elec. Dev.*, (to be published).

[12] R.F. Pierret, "Semiconductor Device Fundamentals," *Addison-Wesley Publishing Company, Inc.*, 1996, pp. 485-492.

[13] K. Shenai, R. W. Dutton, "Current Transport Mechanisms in Atomically Abrupt Metal- Semiconductor Interfaces," *IEEE Trans. Elec. Dev.*, Vol. 35, no. 4, 1988, pp. 468-482.

[14] J. M. Andrews, M. P. Lepselter, "Reverse current-voltage characteristics of metal-silicide Schottky diodes," *Solid-State Electron.*, vol. 13, 1970, pp. 1011-1023.

[15] J. Kedzierski, P. Xuan, E. H. Anderson, J. Bokor, T.-J. King, C. Hu, "Complementary silicide source/drain thin-body MOSFETs for the 20nm gate length regime," *IEDM Tech. Dig.*, 2000, pp. 57-60.

[16] J. Larson, J. Snyder, "Schottky Barrier CMOS," pp. 1-12. Available: http://www.spinnakersemi.com/Documents/SBMOS_Customer_Technical_White paper_Edition3.pdf

[17] B.Y. Tsui, C.P. Lin, "A Novel 25-nm Modified Schottky Barrier FinFET With High Performance," *IEEE Elec. Dev. Lett.*, Vol. 25, no. 6, 2004, pp. 430-432.

[18] Q.T. Zhao, E. Rije, U. Bruer, St. Lenk, S. Mantl, "Tuning of Silicide Schottky Barrier Heights by Segregation of Sulfur Atoms," *Proc. IEEE*, 2004, pp. 456-459.

[19] International Technology Roadmap for Semiconductors (ITRS) (2005), Available: http://public.itrs.net

[20] R. C. Jaeger, "Volume V, Introduction to Microelectronic Fabrication, Modular Series on Solid State Devices," *Addison-Wesley Publishing Company*, 1988, p 58.

[21] J. Kedzierski, D. Boyd, C. Cabral, Jr., P. Ronsheim, S. Zafar, P. M. Kozlowski, J. A. Ott, M. Ieong, "Threshold Voltage Control in NiSi-Gated MOSFETs Through SIIS," *IEEE Trans. Elec. Dev.*, Vol. 52, no. 1, 2005, pp. 39-46.

[22] W.P. Maszara, Z. Krivokapic, P. King, J.-S. Goo, M.-R. Lin, "Transistors with Dual Work Function Metal Gates by Single Full Silicidation (FUSI) of Polysilicon Gates," *IEEE IEDM*, 2002, pp. 367-370.

Chapter 5

Process Modeling Analysis of Bulk Switching SFETs

5.1 SRIM and TRIM Analysis for Implant-to-Silicide (ITS)

As suggested in the analysis in Chapter 4, Section 4.9, the halo dopant concentration, N_{halo}, must be high to generate a lateral field that facilitates sufficient current injection through and over the Schottky barriers at the source/drain M-S junctions. Although said analysis was only performed for V_{DD} = 1.1 V, what becomes clear is that as V_{DD} decreases with each technology node, maintaining the same current density requires a higher N_{halo} and/or a lower SBH to the current carriers in question (assuming the same gate overlap/underlap, *EOT*, and lateral field induced by V_{DS}). To gain more insight into the implant conditions necessary to achieve high N_{halo} values (ignoring the aforementioned solid solubility limitations for post-ITS anneals at ~600 °C), it is useful to perform numerical implant simulations to gain insight into the projected range, ion distribution, and peak ion concentration within a given silicide (namely NiSi, as it is the focus of this study).

A program known as SRIM (Stopping and Range of Ions in Matter) [1] calculates the projected range, longitudinal and lateral straggle, etc. by using a quantum mechanical treatment of collisions between the implanted atoms and the target material. Within SRIM is a program called TRIM (the Transport of Ions in Matter), which can give data on projected range and straggle, but can also give data on ion concentration vs. depth, backscattered ions, etc. for a given implant energy into any type of target, provided the density of the target is known. SRIM and TRIM were used to gain some insight into the

specifics of the ITS process, while Silvaco Athena (SUPREM-IV) was used to try to investigate the details of the post-ITS anneal (shown later in this chapter).

In "making" a NiSi target in SRIM, density is calculated by adding the atomic densities of each element and dividing by the total number of atoms per molecule, which is defined by the stoichiometry (user input). For a NiSi target (with 1:1 stoichiometry), the calculated density is **5.6083 g/cm^3**. However, [2] quotes the molecular density of NiSi to be 4.552x10^{22} "molecules"/cm^3. Dividing this by Avogadro's Number (6.022x10^{23} mol^{-1}) and multiplying by the sum of the molar densities of Si and Ni (28.086 g/mol and 58.69 g/mol, respectively) [1], one finds a different density for NiSi of **6.559 g/cm^3**. Both of these densities have been explored in SRIM and TRIM; the density of 5.6083 g/cm^3 will be referred to as the "low density" case, while the density of 6.559 g/cm^3 will be referred to as the "high density" case.

Figs. 5.1 and 5.2 show, for the low density and high density cases, respectively, the projected range, lateral straggle, and longitudinal straggle of boron and phosphorus into the defined NiSi targets as a function of implant energy from 10-100 keV. While some noticeable difference does indeed exist between the two cases, the results are largely similar, as the densities differ by only about 0.9 g/cm^3. For a near-midgap silicide and without direct gate control (i.e., large *EOT* for a given V_{DD}), or in other words at the body-BOX interface for the SOI and POI devices (where the BOX is 200 nm and 100 nm, respectively), one would expect subsurface leakage to be larger if a halo exists in said region as opposed to a Schottky barrier adjacent to a lightly doped/undoped semiconductor. This would be due to the lower series resistance provided by the halo

region. Thus, unless the body region is very thin, it would seem desirable to confine the halo region near the top of the body region.

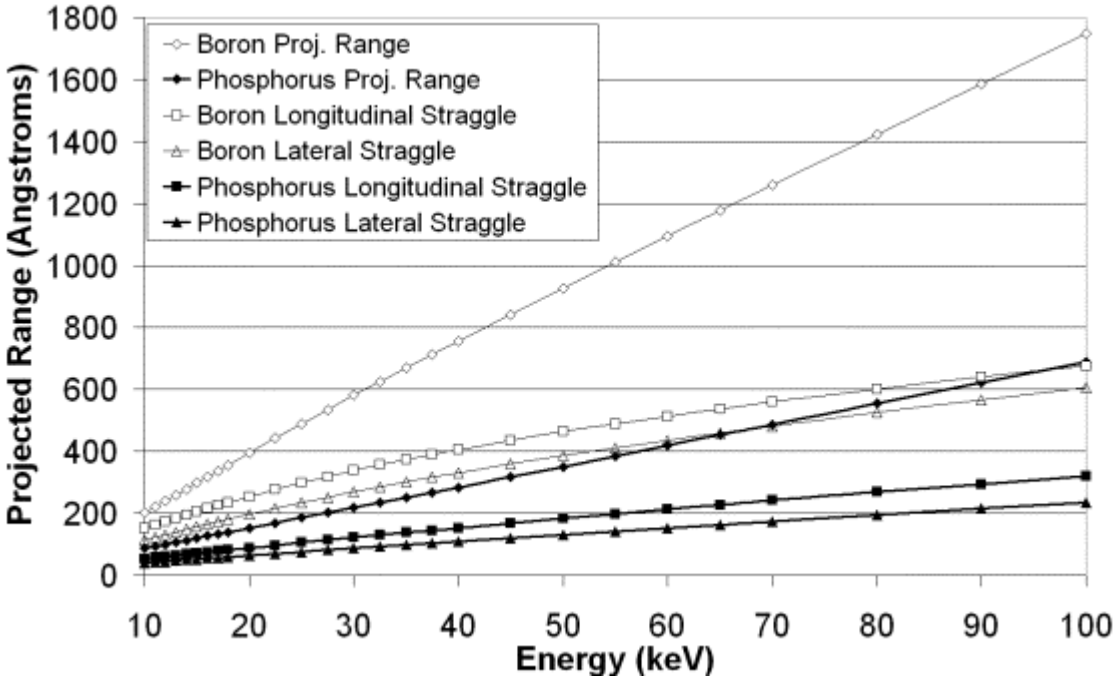

Fig. 5.1. Boron and phosphorus stopping ranges in NiSi vs. energy (low density).

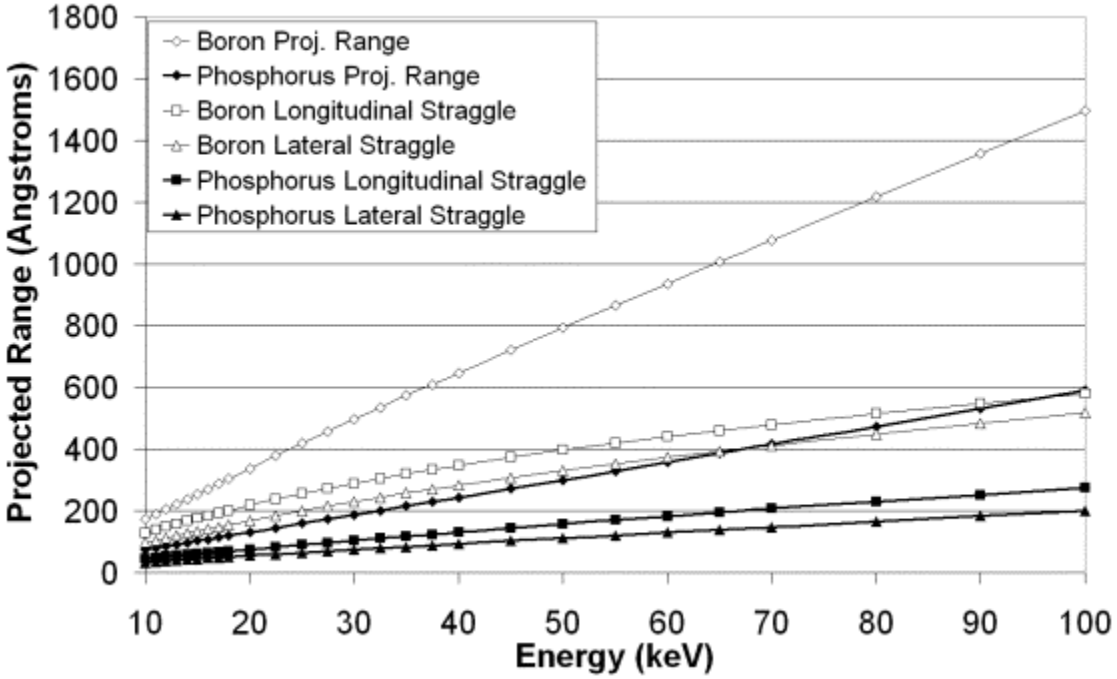

Fig. 5.2. Boron and phosphorus stopping ranges in NiSi vs. implant energy (high density).

Regardless of whether the low density or high density case is considered, either density is considerably larger than the atomic density of silicon (2.3212 g/cm^3) by a factor of about 2.5, which means that the stopping power of NiSi is considerably larger than [crystalline] Si alone (nevermind the added benefit of eliminating crystal damage in the silicon, which now takes place in the silicide), as shown in Fig. 5.3. This not only allows for shallower implants, but also for greater control over the projected range as a function of implant energy – very useful for sub-100 nm processing. This is particularly useful in RIT's SMFL, in which its Varian 350D ion implanter can implant at energies down to 33 keV (lower energies are possible, but this requires some modification and is hence not quite as simple/user-friendly). As such, from Figs. 5.1 and 5.2, for an implant energy of 33 keV, the projected range of phosphorus is about 200-250 Å, while the projected range of boron is about 500-600 Å. Implanting BF_2, however, at 33 keV (which can be treated as a boron implant at 7.4 keV), gives a projected range of less than 200 Å. For body regions of 1000 Å and 2000 Å thicknesses (as is the case for the SOI and POI devices and circuits, respectively), a projected range that is 10-20% of the body thickness would seem reasonable at first glance (more on this later).

Beyond confining dopant atoms near the top of the body region, another implication of Figs. 5.1 and 5.2 is that of the lateral straggle vs. implant energy. Since the device design used in this study uses an oxidation of the polysilicon gate (pre-silicidation) for the sidewall spacer, during the halo implantation, ions may traverse through the spacer and into the body region if the gate is not thick enough (oxide stopping power is similar to silicon). This is of particular concern for ligher dopant atoms such as boron, which can result in a wider halo region than intended if caution is

not taken. However, if the source/drain silicide traverses laterally enough such that it completely underlaps the sidewall spacer (i.e., some gate overlap exists), then the post-implant dopant distribution into the body region is defined primarily by the lateral straggle within the silicide. For a 33 keV boron implant into NiSi in the low density (i.e., worse) case, this lateral straggle is about 300 Å – a convenient match to the target sidewall spacer thickness in this study.

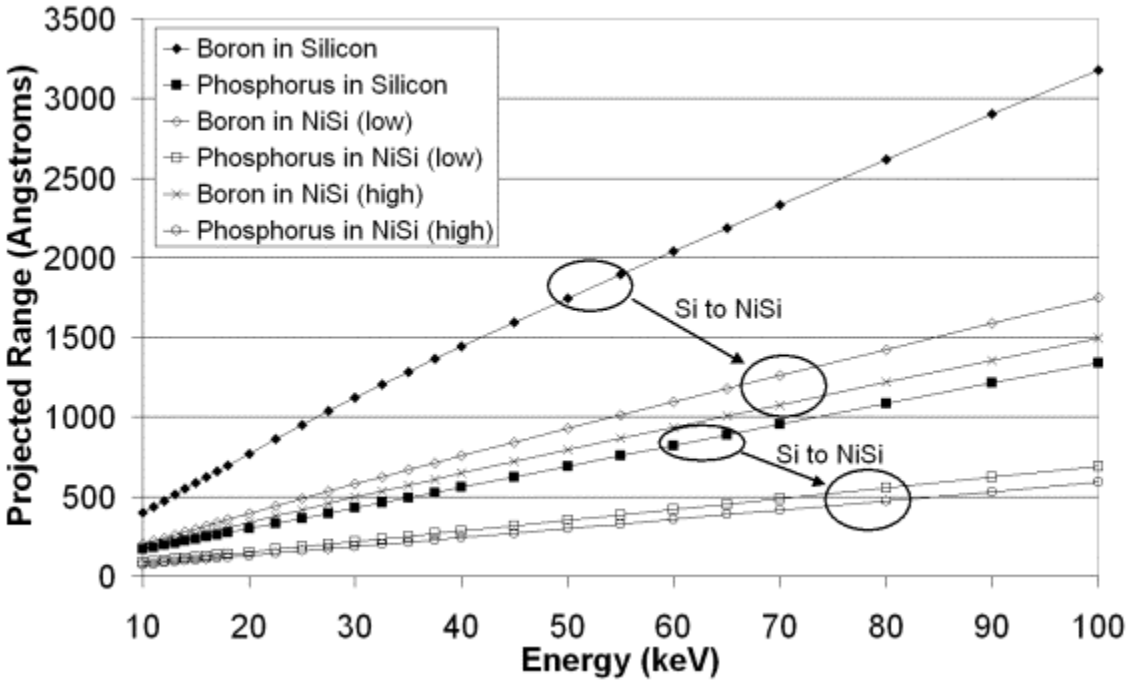

Fig. 5.3. Projected range vs. implant energy, in comparison between silicon and NiSi (both low and high density cases) targets.

Figs. 5.4 and 5.5 show the simulated ion distribution vs. depth into NiSi for both the low density and high density cases, respectively, as predicted from TRIM with a run of 10,000 ions. In both cases, the implant dose was 4×10^{15} cm^{-2}. All implants were performed at 33 keV, and it is interesting to note that the predicted peak ion densities in the NiSi target are considerably larger than the solid solubility limit of each dopant species in crystalline silicon for the subsequent post-ITS anneal that would be performed.

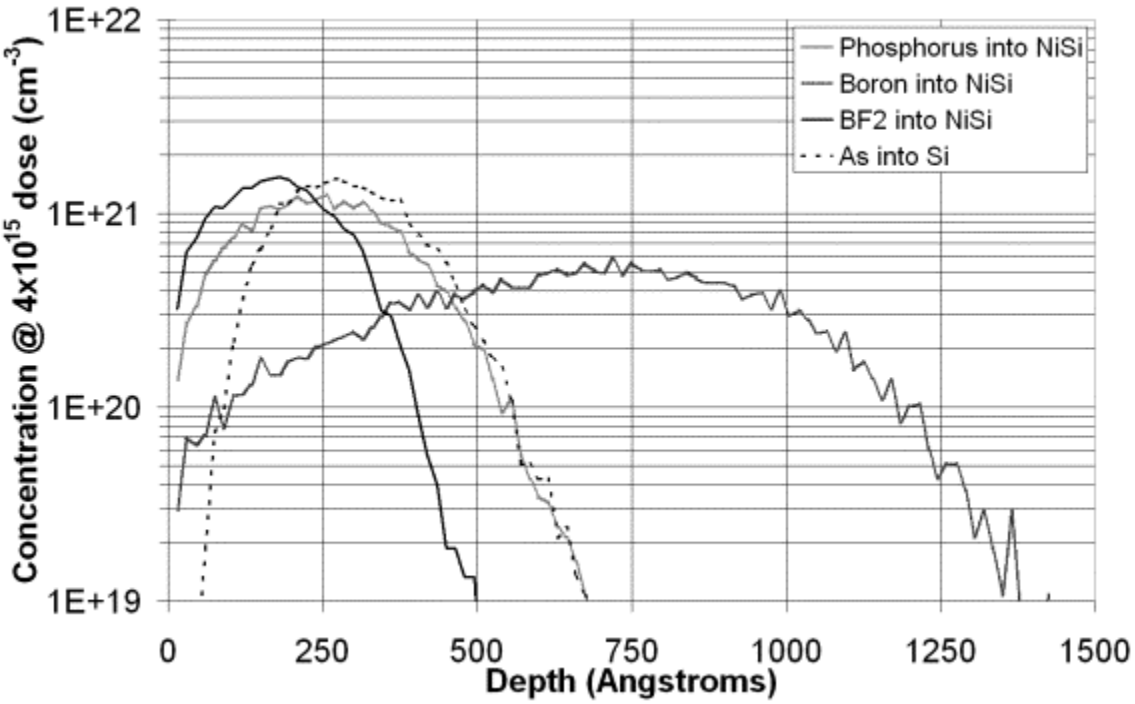

Fig. 5.4. Ion concentration vs. depth into NiSi, compared to As into Si, @ 33 keV, $4x10^{15}$ cm^{-2} dose, low density case, as predicted from TRIM.

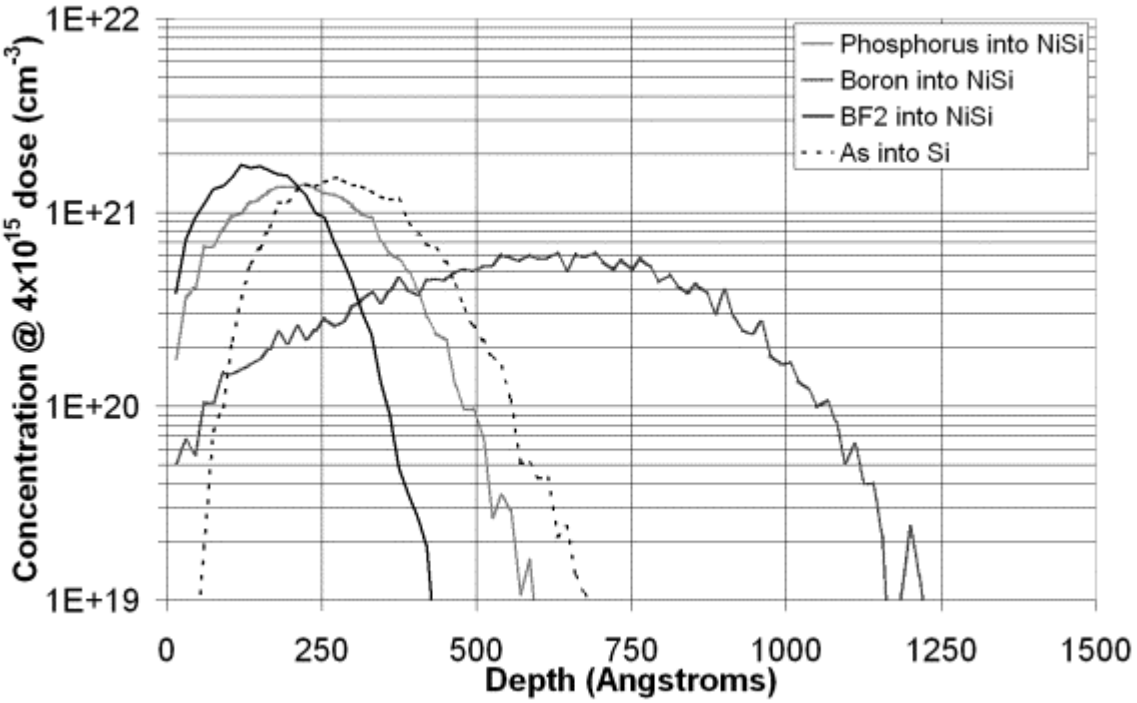

Fig. 5.5. Ion concentration vs. depth into NiSi, compared to As into Si, @ 33 keV, $4x10^{15}$ cm^{-2} dose, high density case, as predicted from TRIM.

For both the low and high density cases, the peak ion concentration in the silicide for phosphorus and BF_2 is greater than $1x10^{21}$ cm^{-3}. Recalling the discussion in Chapter 4, Section 4.9, a simple application of Poisson's Equation to a Schottky diode coupled with the mathematical model developed earlier in Chapter 4 suggests that N_{halo} must be around $7x10^{20}$ cm^{-3} for a maximum current injection of about 2 mA/μm for V_{DD} = 1.1V with NiSi source/drain regions. Figs. 5.4 and 5.5 show that the peak ion concentration in the silicide is considerably higher than the suggested N_{halo} concentration, and so at the very least, the *supply* of the number of ions suggested would indeed exist in the silicide after the appropriate ITS process. While said analysis made certain assumptions that likely result in an underestimation of the actual current density predicted as a function of N_{halo} and SBH, what was made clear is that, for an "ideal" M-S junction, N_{halo} must be highly degenerate. Whether or not this actually happens for the appropriate ITS process remains to be seen, particularly if one considers the solid solubility limitations in silicon at low temperatures.

However, if the supported current injection through and over the Schottky barrier is larger than what the aforementioned theory would predict, then there is indication that the Schottky barrier becomes modified beyond the changes induced by the active dopant-enhanced electric field. There may be an electric field contribution from interstitial (electrically inactive) dopants that lowers the barrier, or the high dopant concentration at *both* sides of the M-S junction may alter the NiSi and Si properties enough to change the branch point, E_B, thus changing the energy that the Fermi level is pinned to (therefore changing the intrinsic barrier height). This would also imply a change in the extent of Heine tail decay, which changes SBL as a function of electric field [3]. In such a case,

the SBH for majority carriers might be lowered to an extent beyond what one would suggest from conventional field-induced lowering, thus facilitating a further increase in the supported current density. This is purely speculative, however, and a complete understanding of such mechanisms extends beyond the scope of the presented work – of primary interest is an experimental understanding of what is gained in Schottky CMOS performance when ITS processes are utilized.

5.2 ITS Modeling with Silvaco Athena

It was mentioned previously that there may be some benefit to confining the implanted dopants near the top of the body region when utilizing an ITS process, as this may potentially support smaller junction depths for the resulting halo region. That is, the halo region may not consume the entire thickness of the body region. However, the diffusivities of dopants in silicides such as $CoSi_2$ and NiSi are orders of magnitude greater than they are in silicon [4], [5]. This suggests that the implanted dopants in the silicide spread out throughout the entire body thickness after a very short period, even at relatively low temperatures. This hypothesis was substantiated after simulation in Silvaco Athena; a 700 °C, 15 sec. post-ITS anneal for a user-defined 1000 Å NiSi film resulted in a uniform dopant spread throughout the entire silicide thickness. Although some discrepancies are surely expected to exist between Athena results for ITS processing and actual results, there seems little reason to think that the enhanced stopping power of NiSi over silicon, in itself, will improve SCE immunity; such is therefore left to the actual *structure* of the device. Additionally, such a result strongly suggests that *the majority of the dopant diffusion during the post-ITS anneal is spent in the silicon region*.

In an attempt to gain some insight into the halo profile at the M-S junction, further simulations in Silvaco Athena were performed. The device structure (Fig. 5.6) was defined manually in Silvaco Athena to minimize the occurrence of grid anomalies during a fabrication sequence that would otherwise be simulated. The gate oxide thickness was set to 100 Å, the BOX thickness was set to 1000 Å, and the source/drain silicide was defined such that the interface to the body region exists with a gate overlap of 20 nm (i.e., with a 30 nm sidewall spacer and a 100 nm body thickness, the silicide diffused 50 nm laterally). Such overlap requires the implanted dopants to diffuse 50 nm before reaching the M-S interface. The body region was defined as n-type with a concentration of $1x10^{15}$ cm^{-3}, and the gate was treated as a FUSI NiSi gate.

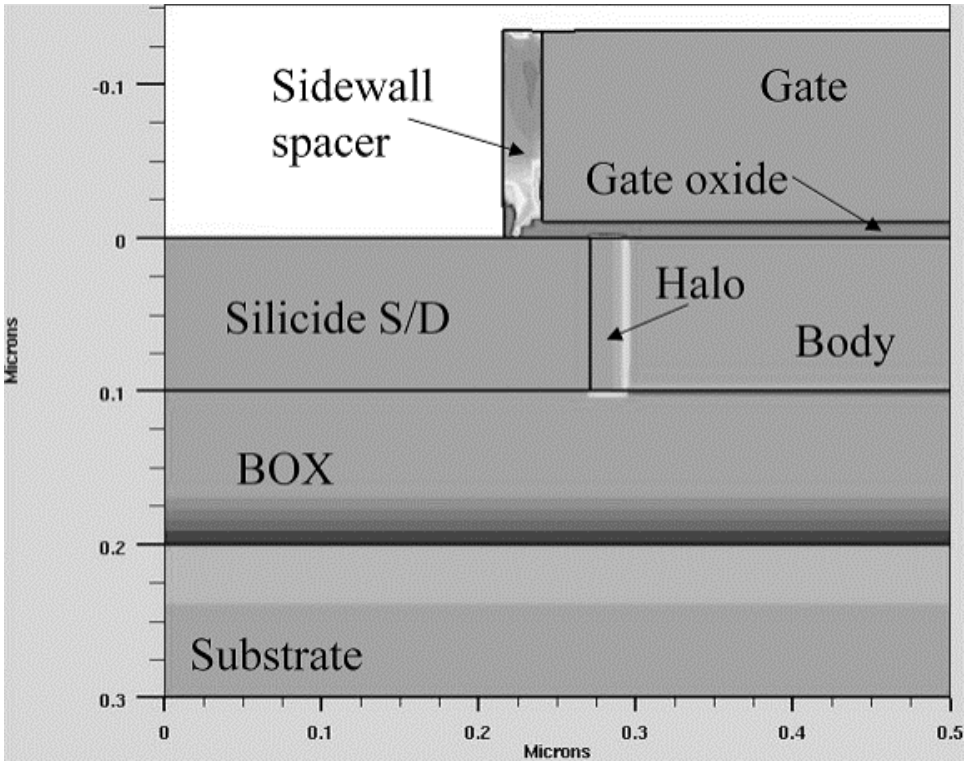

Fig. 5.6. Defined SFET half-structure used in Silvaco Athena, after a 750 °C post-ITS anneal for 30 min. Varying shades of gray represent varying dopant concentrations.

Although NiSi was the chosen silicide for the actual device fabrication, this material is not available in the current version of Silvaco Athena. However, $TiSi_2$ is an available option, and its material properties were redefined to mimic NiSi (6.415 g/cm^3 was used as the density – a mathematical error – although differences from the value of 6.559 g/cm^3 reported earlier in this chapter are negligible). These properties are shown in Table 5.1, although it is noted that the properties related to diffusion within the silicide were kept at the default values for $CoSi_2$ in Athena (Table 5.1) – $TiSi_2$ acts as a poor diffusion source for boron and arsenic due to TiB and TiAs formation, and $CoSi_2$ seems to behave similarly to NiSi for ITS processing [4], [6]. No known data exist regarding the pre-exponential and exponential terms for diffusion of phosphorus and boron within NiSi, although both silicides are known to exhibit high diffusivities for said species.

For the material statement redefining the "tisix" material in Athena		
Parameter	Value	Units
Density	6.415	g/cm^3
Abund.1	1	-
Abund.2	1	-
At.num.1	28	-
At.num.2	14	-
At.mass.1	58.69	amu
At.mass.2	28.086	amu
Diffusion parameters in impurity statements for "tisix" material to mimic NiSi (same values used for both boron and phosphorus)		
Seg.0	1	-
Trn.0	$1.66x10^{-5}$	-
Trn.E	0	eV
Dix.0	4.2	$cm^2/sec.$
Dix.E	2.14	eV

Table 5.1. Parameters for material and impurity statements in Silvaco Athena for initial attempts at simulating ITS processing for NiSi.

The ITS process itself was simulated in Athena by performing a Monte Carlo implant. The post-ITS profile into the "NiSi" material looks similar to those shown in Figs. 5.4 and 5.5, which suggests that the state of the device structure before the diffusion step is in the correct range. Post-ITS anneals were performed at 700 °C and 750 °C (Athena does not seem to be able to effectively simulate diffusion at 600 °C) for both boron and phosphorus after a $4x10^{15}$ cm^{-3} ITS at 33 keV. Although NiSi tends to agglomerate at and above 600 °C, the incorporation of fluorine into the silicide has been shown to increase the thermal stability of NiSi to 750 °C [6]. This higher temperature should allow for higher active dopant concentrations at the M-S interface, thus increasing drive current for a given halo width (particularly relevant for the p-channel device, as boron solid solubility at low temperature is considerably lower than that of phosphorus). The simulated dopant profiles are shown in Figs. 5.7 and 5.8 for 700 °C and 750 °C post-ITS anneals, respectively.

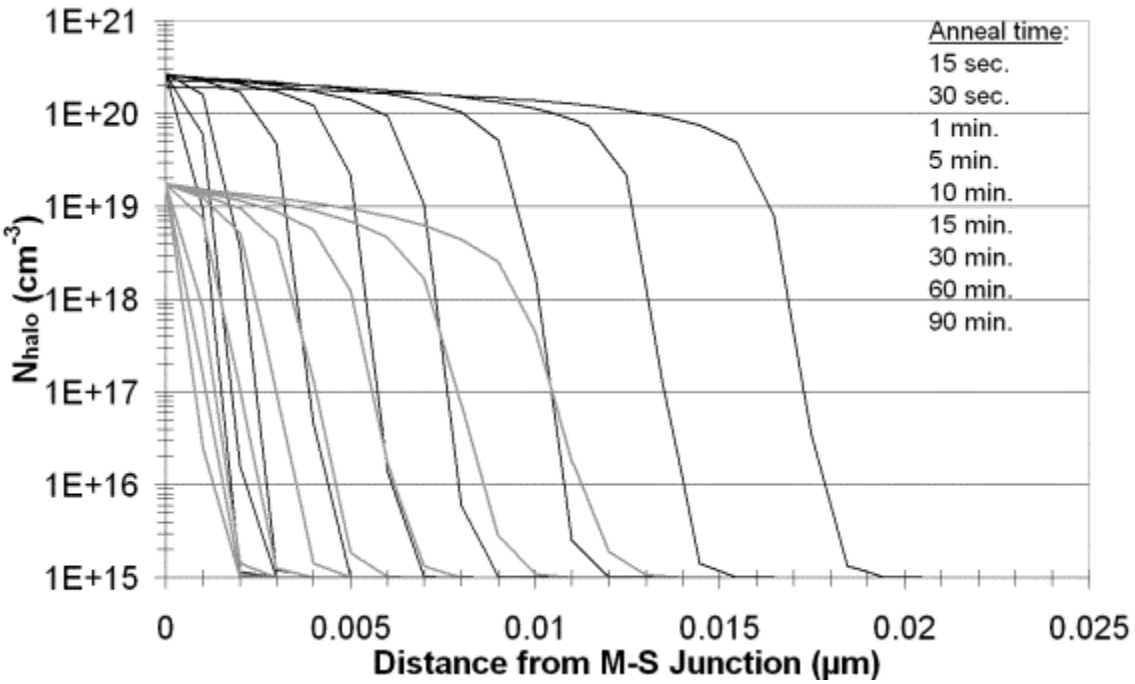

Fig. 5.7. Post-ITS anneal dopant profiles for boron (gray lines) and phosphorus (black lines) at 700 °C for varying anneal times.

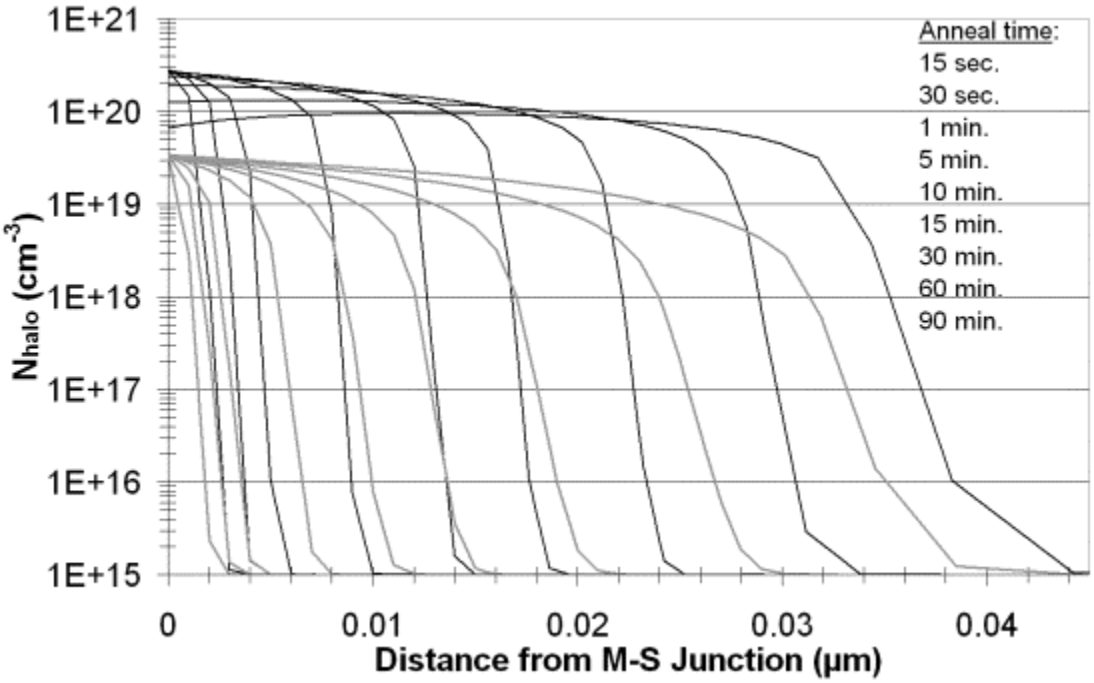

Fig. 5.8. Post-ITS anneal dopant profiles for boron (gray lines) and phosphorus (black lines) at 750 °C for varying anneal times.

It is interesting to observe that the boron dopant profiles in Figs. 5.7 and 5.8 show a smaller halo width than the phosphorus profiles for a given anneal temperature and time. It also seems as if the boron propagation is dramatically underestimated when comparing to experimental data. From [4] and [6], which study ITS for NiSi and $CoSi_2$, respectively, using BF_2^+, the halo width for 600 °C/30 min. and 600 °C/90 min. were found to be 23 nm and 60 nm, respectively, when the implanted dose was kept entirely within the silicide (i.e., no implant propagation past the M-S junction). Fig. 5.8, which is 150 °C higher at 750 °C, does not even come close to those values. At 700 °C for 30 min. and 90 min., the experimental halo widths were 28 nm and 100 nm, respectively, under the same conditions; again, very far off from what is shown in Fig. 5.7 for the boron profile. There are two primary explanations for this. First, both NiSi and $CoSi_2$

are known to enhance boron diffusivity in the adjacent silicon by virtue of vacancy injection [4], [6]. Second, Silvaco Athena does not have a very good diffusion model for boron at relatively low temperatures. One can use the "pls" method statement in the Athena simulation code to more accurately model boron diffusion at low temperatures; however, for some reason this method statement did not allow a diffusion simulation to be performed.

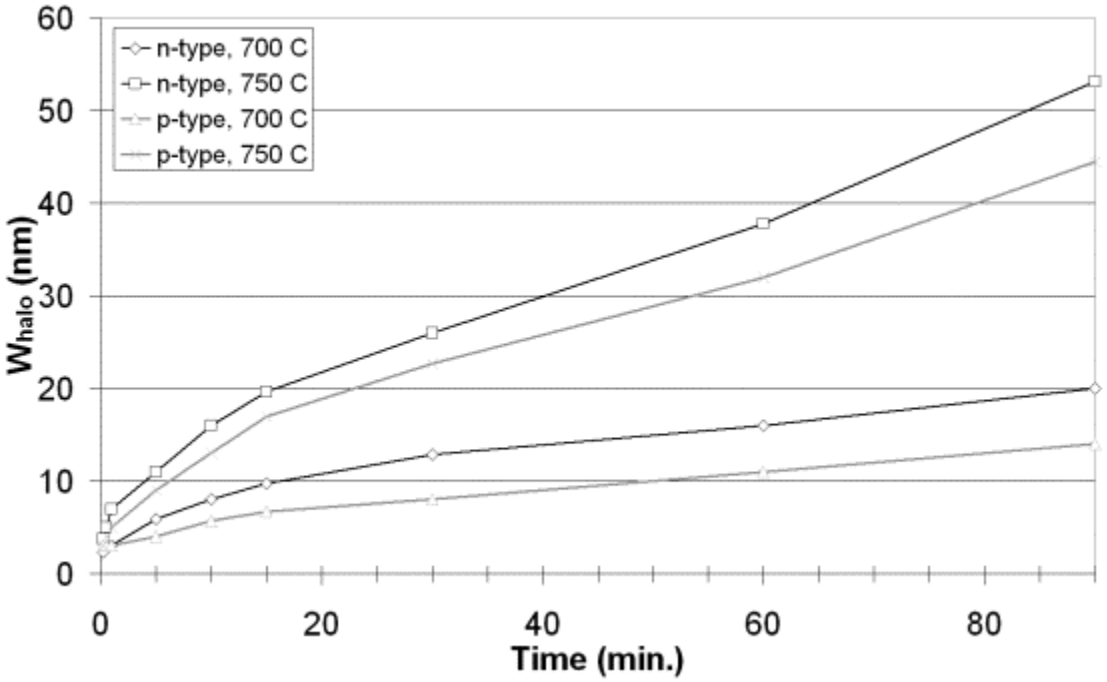

Fig. 5.9. W_{halo} vs. post-ITS anneal time for n- and p-type halos at 700 °C and 750 °C, from Silvaco Athena simulations. W_{halo} is defined where $N_{halo} = 1x10^{15}$ cm^{-3}.

Another important point to note regarding the Athena simulations is that, once the implanted dopants redistribute within the silicide to a uniform concentration, Athena seems to treat the silicide as an infinite source of dopants. In other words, the dopant concentration at any region within the silicide remains constant regardless of how long the simulated diffusion takes place for. As such, the accuracy of the predicted dopant profile is reduced as the post-ITS anneal time is increased to very high values. In spite of

the limitations of this simulation, Fig. 5.9 summarizes Figs. 5.7 and 5.8 in the form of halo width, W_{halo}, vs. post-ITS anneal time.

Although some preliminary numerical modeling was performed in Silvaco Atlas for this device structure, the results will not be shown here due to the primitive and unconvincing nature of the results. To perform "proper" numerical modeling would not only require a self-consistent Poisson-Schrödinger solution for the Schottky barrier, but also an accurate solution to SBL vs. Ψ_s (which Atlas does not currently seem to contain). Without such an SBL solution, the actual current injection through and over the Schottky barrier as a function of N_{halo} and/or V_{GS} is not accurately modeled. Furthermore, the requirement of more accurate solutions to the dopant redistribution within the silicide (i.e., not treating the silicide as an infinite source of dopants) and segregation at the M-S interface confound with the aforementioned requirements to add to the uncertainty of the modeling results. It would seem, then, that a more developed version (i.e., self-consistent Poisson solution, universal mobility, etc.) of the model in Chapter 4 would provide more insight; however, such remains a work in progress which may well extend beyond the scope of this thesis.

5.3 Thermal Budget Implications for NFET and PFET Performance

Although the work presented in [4] noted a post-ITS anneal at 600 °C for 30 min., it has been shown that the thermal stability of NiSi is limited to about 600 °C [6]. This is due to silicide agglomeration and eventual formation of discontinuous islands, whereby the silicide effectively "falls apart." One might therefore contend that the active dopant concentration in the halo region is limited to the solid solubility of the species in question

at 600 °C. However, in [6] it was shown that the introduction of fluorine into NiSi extends its thermal stability up to temperatures of 750 °C. This can help both PFET and NFET drive current; PFETs due to the lower solid solubility limit at a given temperature for boron than phosphorus (i.e., more boron is activated at the M-S interface at 750 °C vs. 600 °C), and NFETs due to the lower amount of diffusion of phosphorus than boron for the same thermal process (i.e., more phosphorus diffusing toward the M-S interface at 750 °C vs. 600 °C).

For boron implants, such an implementation is as simple as implanting BF_2 instead of B_{11}, which, as was shown in Figs. 5.4 and 5.5, has the added benefit of reducing the projected range and straggle. N-type dopants, however, are not coupled with fluorine, and so the thermal budget of a Schottky CMOS process using NiSi ends up being limited by the NFETs. In light of the findings in [6], however, it would seem that this thermal budget limitation can be mitigated by performing dual implants for the NFETs – one fluorine implant and one phosphorus (and/or arsenic) implant.

Fig. 5.10 illustrates the phosphorus and fluorine profiles (Fig. 5.11 shows BF_2 and fluorine profiles) into NiSi predicted by TRIM in the aforementioned low density and high density cases for an implant energy of 33 keV and a dose of $4x10^{15}$ cm^{-2}. As fluorine is a lighter species than phosphorus, its distribution throughout the source/drain depth is wider and the peak concentration is lower. For a structure where the implanted dopants are confined toward the top of the silicide (i.e, a relatively thick body), such as the work presented here, the fluorine spread throughout the entire silicide thickness should prevent/minimize agglomeration of the entire silicide, particularly in the subsurface region where the gate has less control over leakage.

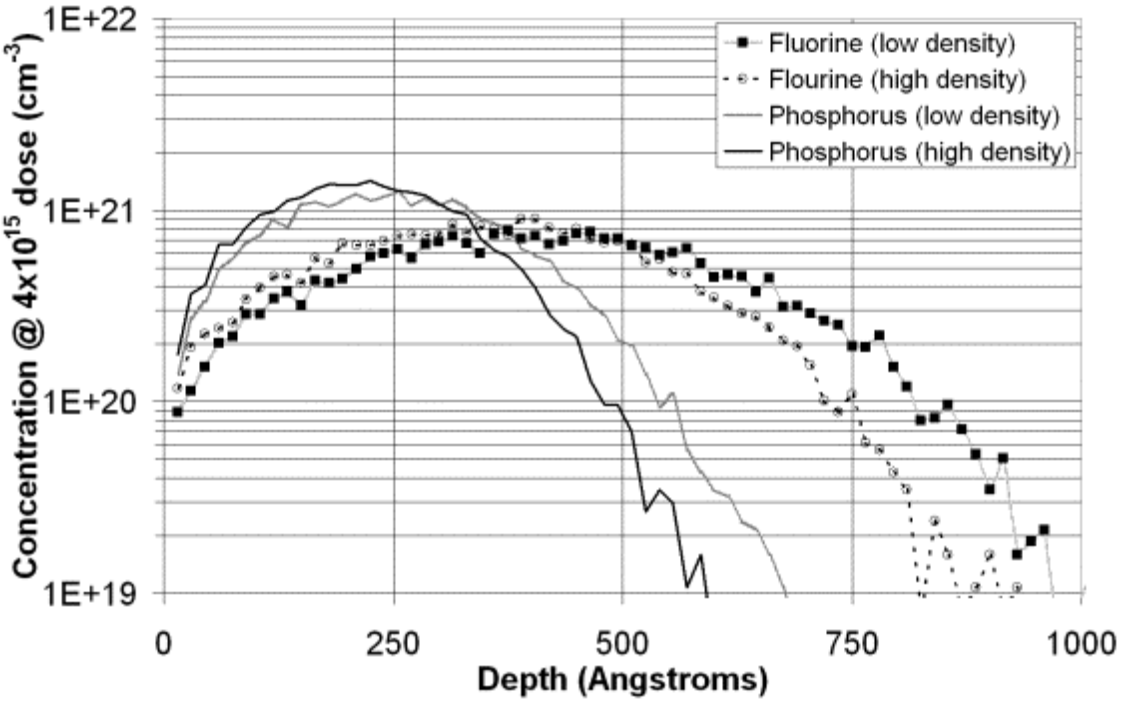

Fig. 5.10. Ion concentration vs. depth into NiSi for phosphorus and fluorine implants (both $4x10^{15}$ cm^{-3} @ 33 keV), as predicted from TRIM.

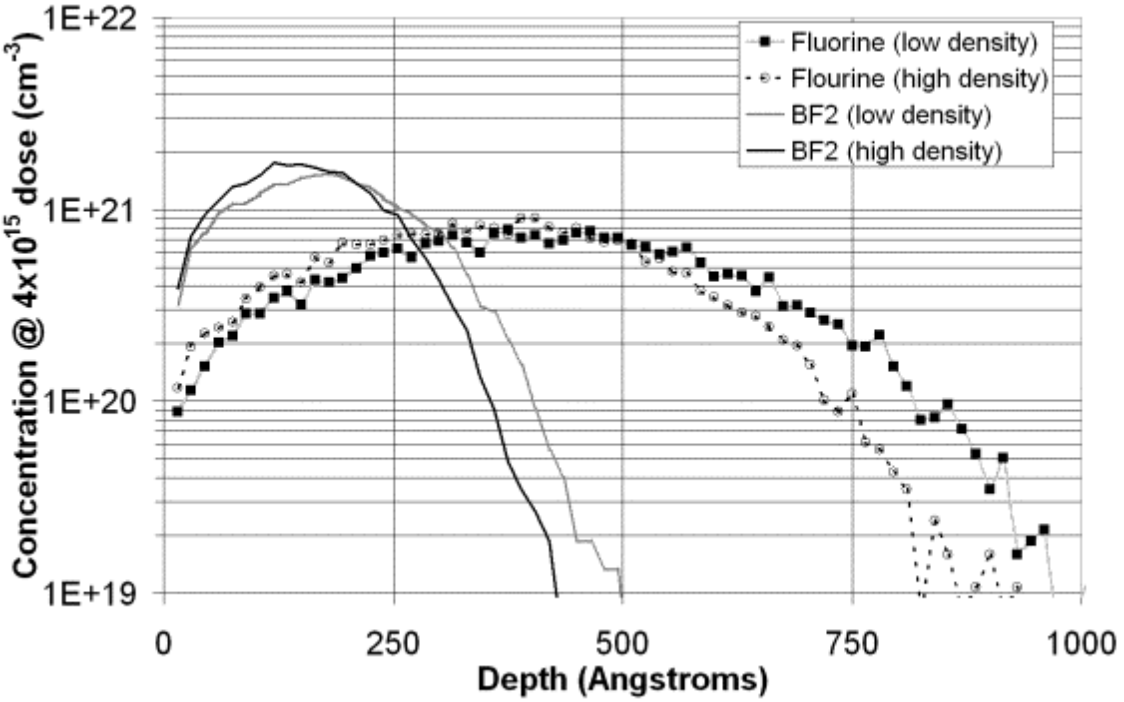

Fig. 5.11. Ion concentration vs. depth into NiSi for BF_2 and fluorine implants (both $4x10^{15}$ cm^{-3} @ 33 keV), as predicted from TRIM.

A final point to note regarding fluorine implantation is that it can be used to "preamorphize" the NiSi before subsequent implantation with phosphorus/arsenic and BF_2. The induced damage within the silicide (considered to have a higher dependence on dose than energy [4]) decreases the grain size. Since dopant diffusion within the silicide has been attributed to grain boundary diffusion [2], it would not be unreasonable to suspect enhanced dopant diffusivity within the silicide (and perhaps segregation at the M-S junction), thus potentially increasing the dopant concentration at the interface. This would be more beneficial for the NFETs, as the larger phosphorus/arsenic atoms tend to diffuse slower than boron. Although the increased damage from the fluorine implant will increase the sheet resistance of the silicide, this cost must be weighed against the potential benefit of achieving higher current injection at the M-S junction. This in itself is a subject worthy of diligent study.

Chapter 5 References

[1] J.F. Ziegler, J.P. Biersack, Stopping and Range of Ions in Matter, Available: http://www.srim.org

[2] K. Maex, M. Van Rossum, "Properties of Metal Silicides," *Short Run Press, Ltd.*, 1995, p 20, 298-306.

[3] K. Shenai, R. W. Dutton, "Current Transport Mechanisms in Atomically Abrupt Metal-Semiconductor Interfaces," *IEEE Trans. Elec. Dev.*, Vol. 35, no. 4, 1988, pp. 468-482.

[4] B.-S. Chen, M.-C. Chen, "Formation of cobalt-silicided p^+n junctions using implant through silicide technology," *J. Appl. Phys.*, Vol. 72, no. 10, 1992, pp. 4619-4626.

[5] K. Maex, M. Van Rossum, "Properties of Metal Silicides," *Short Run Press, Ltd.*, 1995, pp. 298- 306.

[6] C.-C. Wang, C.-J. Lin, M.-C. Chen, "Formation of NiSi-Silicided p^+n Shallow Junctions Using Implant-Through-Silicide and Low-Temperature Furnace Annealing," *J. Electrochem. Soc.*, Vol. 150, no. 9, 2003, pp. 557-562.

Chapter 6

Device and Circuit Design

6.1 Limitations for Schottky CMOS on Bulk Substrates

It was the initial intent of this work to demonstrate Schottky CMOS using bulk substrates as opposed to SOI substrates due to the lower starting cost of bulk substrates; however, it was later realized that doing such with a *single* metal silicide and the approach discussed would result in a longer development time, lower circuit density, potentially lower yield, poorer performance, and ultimately a higher total cost. The cross-section for this initial design for an inverter is shown in Fig. 6.1, where the pull-up and pull-down networks have body regions of the same dopant type (n-type in the case of Fig. 6.1). The pull-up network would be a conventional SFET, while the pull-down network would be a bulk switching SFET, both separated by some form of trench isolation, and a negative V_{DD} ($-V_{SS}$) would be utilized in the case of n-type body regions. The purpose of the oppositely doped well was to prevent current flow between the V_{SS} and ground terminals of the inverter, as well as substrate leakage.

The idea behind the design in Fig. 6.1 was that the pull-up network would modulate tunneling current through the Schottky source/drain regions, while the pull-down network would modulate thermal current over the halo-body thermal barrier. However, such a design was proposed before the modeling efforts from Chapter 4 were performed, and so the ineffectiveness of the proposed design was not yet fully realized, as the dominance or lack thereof of tunneling current in SFET operation, and hence the performance of the proposed pull-up network in Fig. 6.1, was not yet apparent. An

understanding of the dominant current mechanisms in SFETs of various designs does indeed have substantial implications for device and circuit design, as was ultimately realized in the aforementioned mathematical endeavor.

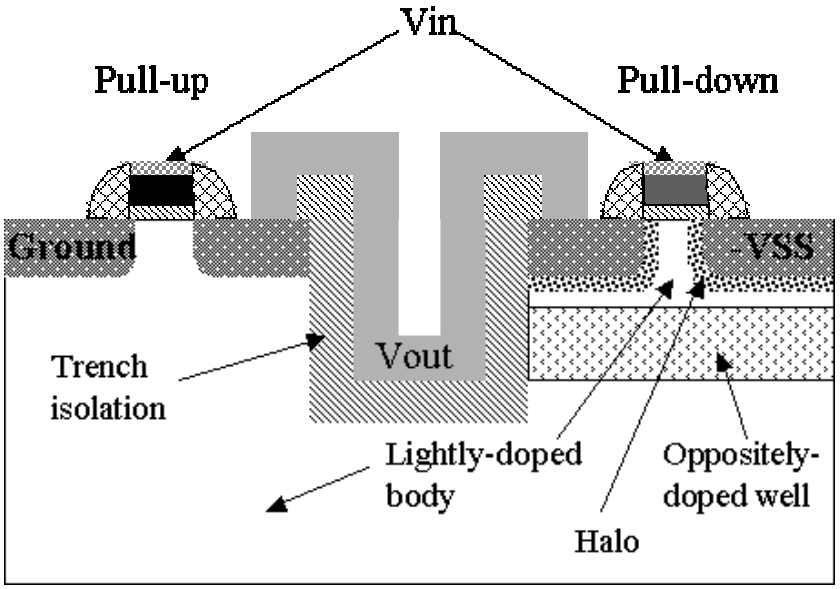

Fig. 6.1. Cross-section of initial Schottky CMOS inverter design proposed for bulk substrates.

Certainly, bulk switching pull-up *and* pull-down networks could be demonstrated on bulk substrates, but this raises process complexity issues beyond those in Fig. 6.1, as an additional two implant steps would be required (well and halo implants for the pull-up network). Also, the circuit density potential of this technology would not be achieved due to the requirement of trench or LOCOS isolation between the NFETs and PFETs. An additional, but beneficial, detail of such an approach is that the V_{DD} and ground terminals in Fig. 6.1 would be reversed, as dual bulk switching networks (of opposite doping) enable the use of positive V_{DD} values.

In considering non-bulk switching (i.e., conventional) SFETs on bulk silicon for both the pull-up and pull-down networks, for the moment the key limitation is a material

constraint, as the currently known materials that would provide the best performance for such an implementation are fairly expensive. As mentioned in Chapter 3, platinum is currently the most promising material for conventional p-channel SFETs; however, in a high-volume manufacturing environment where several thousands of wafers are started every week, the material cost of platinum deposition targets alone may be large, potentially driving up the cost of the end product. Also, while in Chapter 4 it was shown that PtSi, when used properly, can result in acceptable performance for 25 nm p-channel SFETs, the requirement for moderate to high body doping levels in a bulk silicon implementation is not optimal in a device size regime where discrete dopant effects play a considerable role in performance variation. Regardless, a conventional Schottky CMOS implementation on bulk silicon has been proposed by Tucker [1], and is illustrated in Fig. 6.2. In said implementation, PtSi and $ErSi_2$ are used as the PFET and NFET source/drain regions, respectively, while $CoSi_2$ is used as inter-device isolation and as a local interconnect. As will be shown in Section 6.2, a similar approach can be pursued on SOI substrates, but with much enhanced simplicity and circuit density.

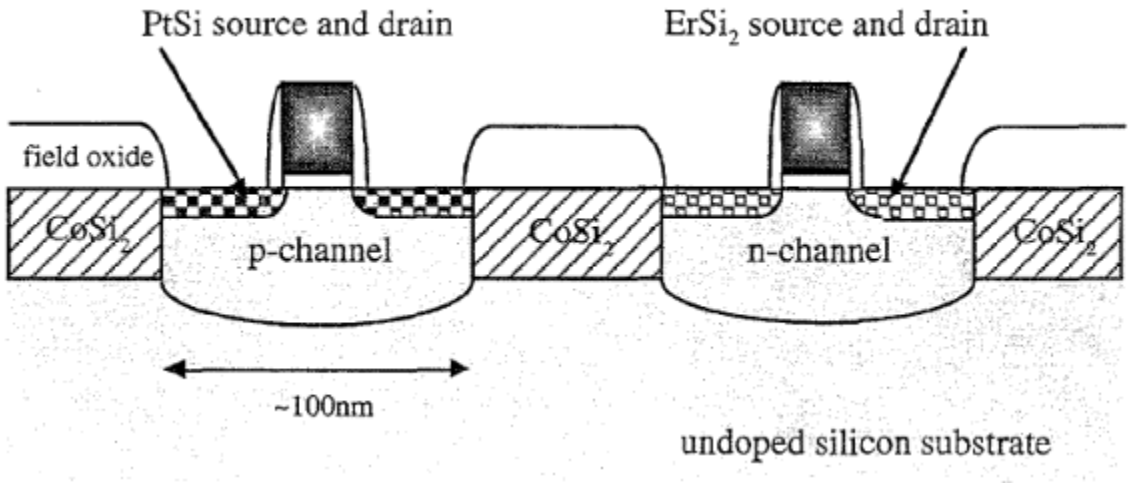

Fig. 6.2. Illustration of Schottky CMOS inverter using conventional SFETs on bulk silicon, adapted from [1].

Regardless of the approach taken to demonstrate Schottky CMOS on bulk substrates, the most inherent disadvantage over SOI substrates is circuit density, as even for an inverter, the pull-up and pull-down networks must be isolated from each other by some sort of trench or material. As such, the only gain in switching from conventional CMOS to Schottky CMOS on bulk substrates, besides simpler fabrication, would be whatever intrinsic gain in device performance that can be achieved. By virtue of SFETs using metallic source/drain regions, however, there is potential for these regions to serve multiple purposes when the appropriate substrate is utilized, thus effecting some level of change in design methodology.

6.2 Device and Circuit Architecture for SOI Substrates

For any type of SOI device, the body region has a finite thickness, whereby the lower boundary is defined by the interface to the buried oxide (BOX). For conventional CMOS on SOI substrates, the pull-up and pull-down networks are isolated from each other by this BOX and by shallow trench isolation (STI), which is filled with oxide. Thus, the devices are entirely isolated from each other by some layer of oxide, as Fig. 6.3 illustrates. As a result of this, the gate pitch must be increased due to the tighter via pitch, as the V_{out} terminal must run over the inter-device isolation to connect the source/drain region of the NFET to the source/drain region of the PFET. While for conventional CMOS there is a potential layout advantage with SOI over bulk substrates, due to the elimination of the body contacts, the continued existence of the inter-device isolation, although smaller, leaves room for improvement.

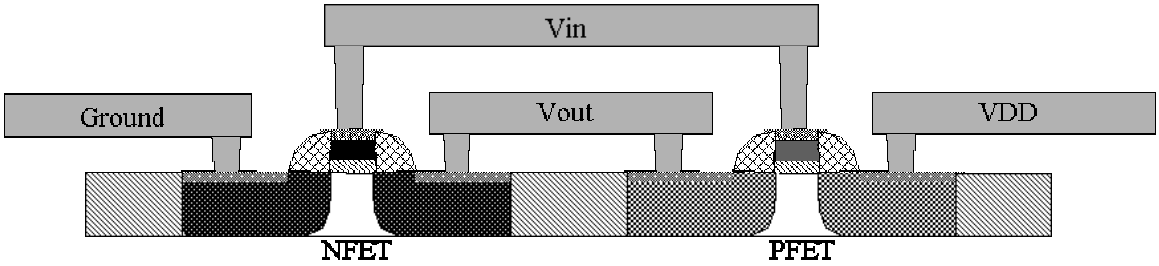

Fig. 6.3. Illustration of a conventional CMOS inverter on SOI substrates. The BOX is not shown, but would be placed under the active region.

If the source/drain material for both the NFET and PFET are metallic, however, and if this material is the same material for both devices, then the V_{out} terminal can be shared between both devices (roughly doubling as a first level interconnect), as would be the case for single metal Schottky CMOS on SOI substrates. The key advantage here is that the body region is fully silicided, which, on top of device performance advantages, allows for the source/drain regions to also act as inter-device isolation, as well as some inter-circuit isolation (some STI would still be required). In doing so, the silicide serves *four* functions – source/drain region for the NFET, source/drain region for the PFET, inter-device and some inter-circuit isolation, and the first level interconnect. Additionally, the gate pitch is not limited by the via pitch in such a structure. Such an implementation is illustrated in Fig. 6.4, where the NFET and PFET are n-channel and p-channel bulk switching SFETs, respectively.

By using the source/drain region as inter-device isolation, the pull-up and pull-down networks (PFETs and NFETs, respectively) can be placed closer together without having to decrease the device size, since the V_{out} terminal is shared between the NFET and PFET (it only needs to be wide enough for one via to contact said region). For present and future technology nodes, there should not be a lithographic process constraint

to doing this, as current devices have gate lengths that are already smaller than the half pitch of the lithography processes in use. While placing the devices closer together like this may ultimately increase the gate length by some increment due to the reduced image quality resulting from the reduced pitch (though the extent of this can be mitigated by resist trimming or other pattern formation "tricks"), the *circuit* is smaller, and so the overall performance should increase. In addition, since the pull-up and pull-down networks can share a source/drain region (which, again, implies only one via contact to the V_{out} terminal of an inverter instead of two separate contacts), the local interconnect density can be maintained or even reduced as the circuit density is increased with this approach. In other words, the "first level interconnect" function that the metallic source/drain regions serve can be utilized to displace a considerable amount of local via contacts and interconnects. This effect is perhaps more substantial than the intrinsic increase in circuit density and device performance, as interconnect density tends to limit circuit size, while interconnect delay tends to limit circuit performance at very aggressive scales.

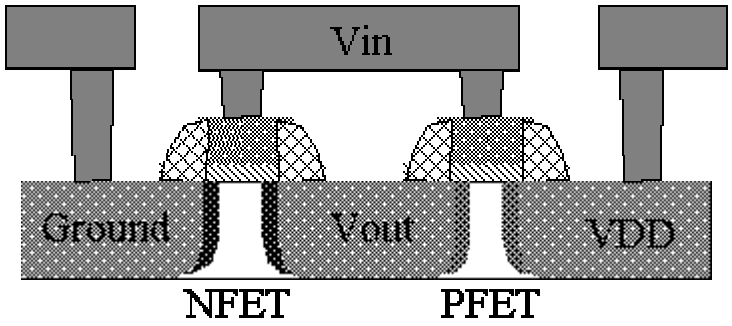

Fig. 6.4. Illustration of single metal Schottky CMOS inverter on SOI using bulk switching SFETs. The V_{out} contact to the first metal level would be located in the third dimension, into or out of the page. The BOX is not shown, but would be placed under the active region.

From another perspective, such an implementation can allow for larger devices at smaller technology nodes. For example, if the circuit density gain is on the order of 50 % (roughly the gain achieved from moving to the next technology node, such as 90 nm to 65 nm), then devices of the same size can be used for the next node as a "low power" solution, due to the reduced leakage offered by the larger channel length, without sacrificing circuit density. Effectively, then, such an implementation has the potential to create a new technology node where one did not exist previously, thus extending the "life" of silicon-based CMOS scaling by another 2-3 years.

The Schottky CMOS approach taken in the work presented is very similar to the illustration in Fig. 6.4, with the only significant difference being the exclusion of via contacts to the gates and source/drain regions. STI between circuits was excluded, thus resulting in MESA isolation, and the BOX was treated as a "pseudo-ILD," upon which the Metal 1 and gate material would lie. Thus, the Metal 1 lines simply overlapped/ran over the silicide source/drain regions and the gate regions. This was done to keep the process as simple as possible, as further complications tend to run the risk of reducing yield. The source/drain regions were fabricated using NiSi, and an ITS process was used to define the NFET and PFET halo regions. An example of this circuit layout is illustrated in Fig. 6.5.

Since no via contacts were used to connect to the active region, the V_{out} portion of the active region between the NFET and PFET gates was extended, as Fig. 6.5 shows, to allow for a Metal 1 contact region of sufficient size without seriously affecting the total area consumption of the "main body" of the inverter (the 1 μm x 5 μm active area rectangle between the Ground and V_{DD} contacts). Naturally, the use of via contacts

would facilitate a reduction in the surface area consumption of this inverter – an optimized inverter design for 0.5 μm devices and a 1:1 NFET:PFET width ratio with W_{ch} = 1 μm would be on the order of 3 μm vertical by 4 μm horizontal, or 12 μm^2.

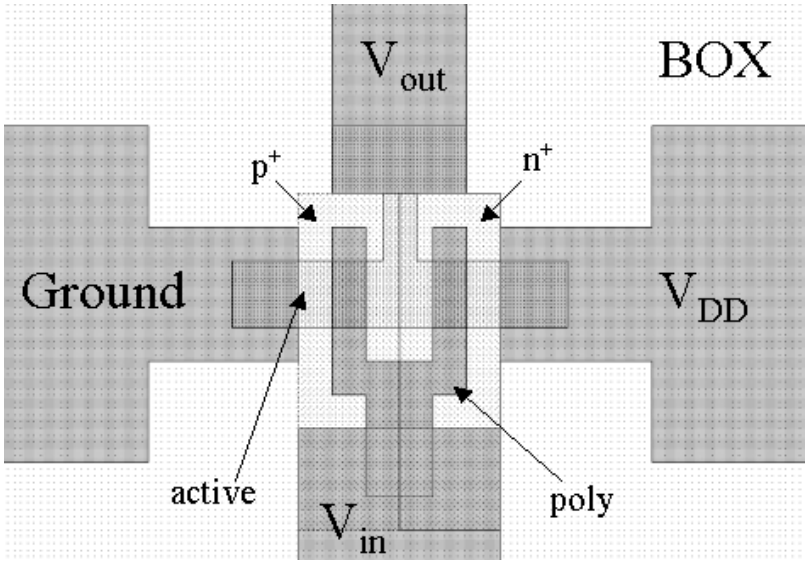

Fig. 6.5. Schottky CMOS inverter layout example. Mask-defined L_{ch} = 0.5 μm, PFET-to-NFET width ratio = 1:1, channel width (W_{ch}) = 1 μm.

In the layout for an SRAM cell, it would be possible to achieve sub-20 μm^2 area consumption for 0.5 μm devices using single metal Schottky CMOS on SOI substrates (assuming via contacts and one or two metal levels), as Fig. 6.6 shows. Assuming a 50 % surface area reduction per node, this size would be on the order of what is expected on bulk substrates using 0.35 μm devices [2], if not significantly below what has been achieved with 0.5 μm devices [3] – [6]. Although the design rules used in Fig. 6.6 are somewhat aggressive, SRAM design usually employs aggressive layout rules compared to the rest of a given microchip so as to maximize density – process latitude is not the top priority. At least for the SRAM cell, then, a circuit density gain on the order of an entire

process generation is indeed possible by simply switching from conventional CMOS to single metal Schottky CMOS using SOI substrates with devices of the same exact size. It should be noted that SRAM cells were not fabricated in this study, due to the use of only one metal level and no via contacts, and so the illustration in Fig. 6.6 is purely conceptual.

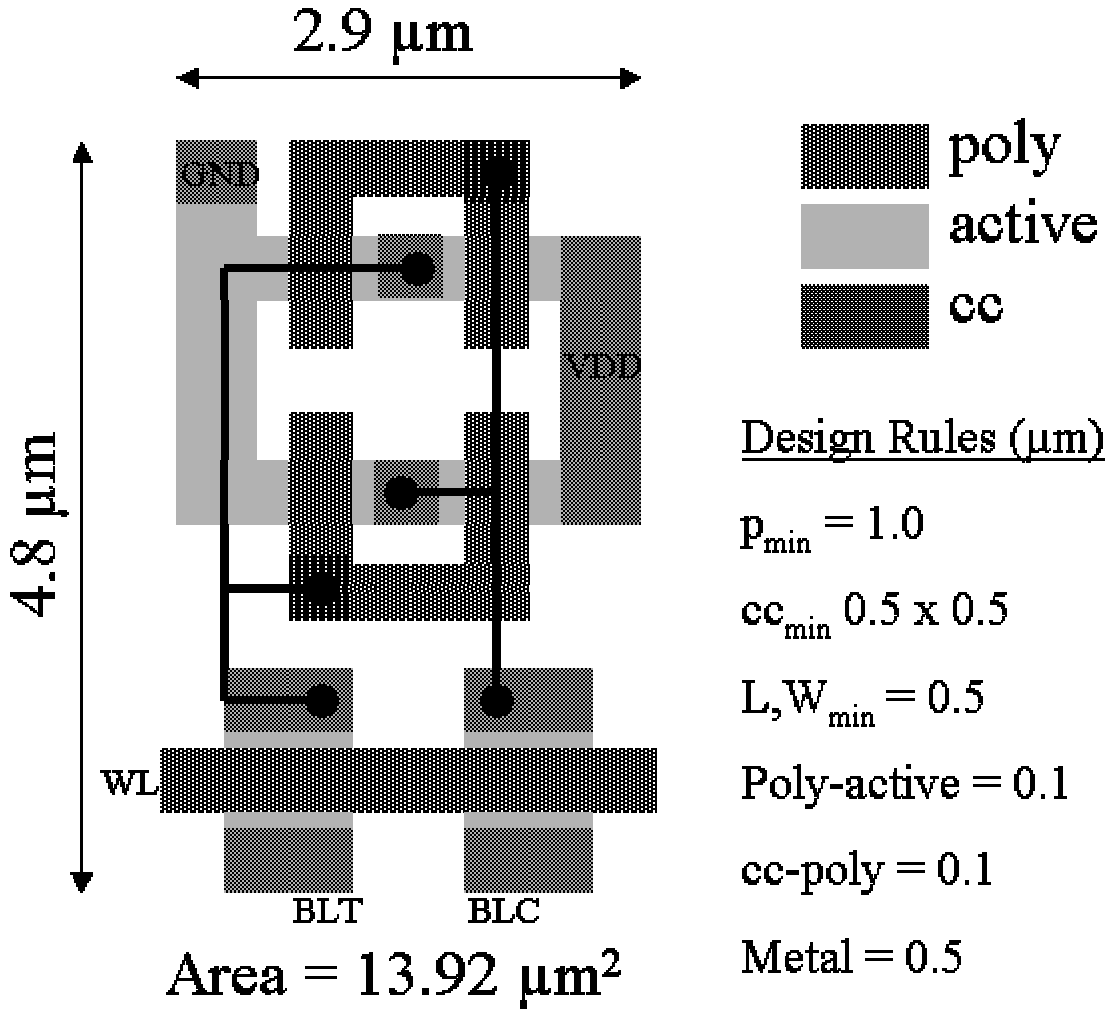

Fig. 6.6. SRAM cell illustration using single metal Schottky CMOS on SOI substrates. The metal lines were not drawn to their full width to give more visibility to underlying layers.

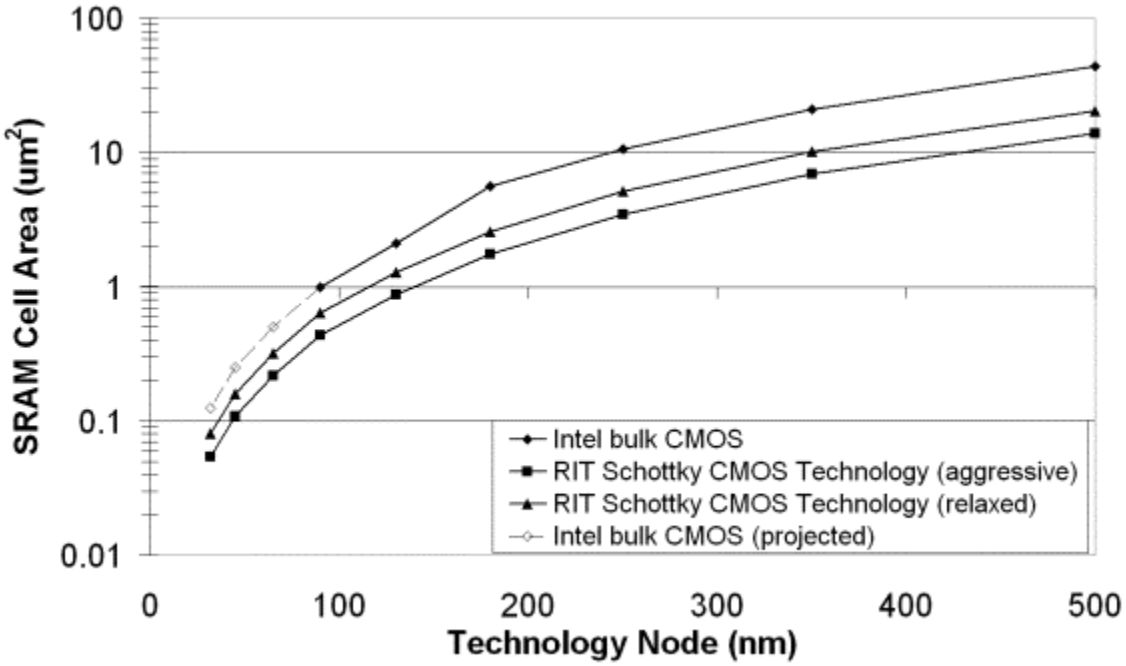

Fig. 6.7. Comparison of single metal Schottky CMOS technology with aggressive and relaxed design rules to Intel's bulk CMOS technology with regard to SRAM circuit density per technology node.

To further quantify the potential circuit density advantage, in extrapolating to the expected circuit density from Fig. 6.6 for more modern process technologies by assuming a 50% density increase per technology node, one obtains the curves in Fig. 6.7. Both aggressive (Fig. 6.6) and relaxed design rules (poly-to-active and cc-to-poly from Fig. 6.6 are 0.25 μm rather than 0.1 μm) are compared to Intel's bulk CMOS SRAM density [7] down to the 90 nm node. Even for the relaxed design rules, a significant increase in density is achievable – on the order of one technology node, as expected. The aggressive design rules achieve an increase in density equivalent to almost *two* entire nodes for a given device size. It is noted that such is perhaps not a "fair" comparison, though, as bulk technology is compared to SOI technology, but it does give some insight into what is possible compared to mainstream technology. For example, at the 180 nm node,

Intel's bulk CMOS SRAM technology fits one cell within a 5.6 μm^2 area. The extrapolated surface area consumptions for single metal Schottky CMOS at this node for the relaxed and aggressive design rules are 2.56 μm^2 and 1.74 μm^2, respectively, assuming no change in circuit layout/design compared to that in Fig. 6.6.

Although it has been suggested here that the potential of single metal Schottky CMOS is quite promising, at least with regard to circuit density, a primary challenge from an integration perspective is ensuring full silicidation of the source/drain regions without any excess silicidation. In the case of nickel, the ratio of nickel thickness to consumed silicon thickness to NiSi thickness is 1:1.84:2.22 [8]. In other words, a given thickness of nickel consumes 1.84 times its thickness in silicon to form 2.22 times its thickness in the form of NiSi [8]. For an SOI wafer with a 100 nm body thickness, such as the substrates used in this study, the silicon in the source/drain regions will be on the order of 85 nm after growing a 10 nm gate oxide and a 30 nm oxide sidewall spacer (discussed later). This requires the deposited nickel thickness to be about 46.2 nm for full silicidation. If the nickel is thicker than this, excessive lateral diffusion will take place, as will the onset of void formation within the body region, thus reducing device performance [8], [9]. If the nickel is not thick enough for full silicidation, then the V_{out} terminal cannot act effectively as intra-device isolation due to the existing leakage path between the BOX and the silicide source/drain regions, thus reducing circuit performance. *Extreme precision* with the nickel deposition is therefore necessary to achieve optimal device and circuit performance for this particular implementation of single metal Schottky CMOS. As such, at highly aggressive scales, it is perhaps the

quality of this metal deposition step (and subsequent silicidation [10]) that must receive the highest priority.

Reverting back to the device architecture used in this study, for both the SOI and POI substrates, the body doping for both the NFETs and PFETs were kept as low as possible. This was simple for the POI substrates, because the as-deposited polysilicon is undoped. For the SOI substrates, the starting material was lightly doped p-type (around $1x10^{14} – 1x10^{15}$ cm^{-3}), with a body thickness of 100 nm. For the NFETs, therefore, the body was counterdoped to achieve an n-type dopant concentration within the same range. While this counterdoping decreases the channel mobility for the SOI NFETs, the total dopant concentration is still fairly low, and so the room temperature electron mobility remains at around 1300 $cm^2/V{\cdot}sec$ [11]. Likewise, due to the low PFET body doping, the room temperature hole mobility is about 460 $cm^2/V{\cdot}sec$ for the SOI PFETs. The POI wafers experienced a high temperature anneal after deposition (1 hour at 1100°C), and so the electron and hole mobilities for the POI SFETs are unknown due to the unknown dependence of mobility on grain size. It is presumed, though, that the grain size is on the order of 0.2 μm (the thickness of the POI body regions), as conventional thermal annealing of polysilicon and amorphous silicon films tends to be columnar.

Another important feature of the devices fabricated in this study is the attempt to form a fully-silicided (FUSI) gate simultaneously with FUSI source/drain regions. The purpose for this was two-fold. First, a FUSI gate would eliminate polysilicon depletion, thus reducing *EOT* whereby its only component is the gate oxide thickness (100 Å target for the devices in this study), which is independent of gate and drain bias. Second, as NiSi was used as the silicide for these devices, NiSi FUSI gates give a roughly midgap

characteristic (although slightly on the p-type side, by about 0.1 V). As bulk-switching SFETs are not entirely "off" when zero gate bias is applied due to the nature of the body region, it was the intent to utilize the near-midgap nature of NiSi to pull the NFET and PFET transfer characteristics as close to the off state as possible with minimal process complexity (i.e., to form a gate-induced thermal barrier rather than a dopant-induced thermal barrier). In other words, while conventional CMOS uses a p+ poly gate for the PFETs (which would have an n-type body region) and vice versa, for the approach taken in this study, the PFET body region is of a different type (intrinsic or p-type) and so would instead need an n+ poly gate (vice versa for the NFET) to achieve an acceptable off state. Rather than perform separate poly and halo implants, however, the gate was kept to a single type for both the NFETs and PFETs, which, as an almost-midgap gate, should provide a good compromise for optimal NFET and PFET gate workfunctions. Also, as the lateral halo dopant propagation during the post-ITS anneal was expected to be very small (less than 10 nm) [12], it was presumed that, on the scale of ~0.5 μm+ device sizes, the percentage-wise reduction in the channel region that constitutes the halo region would compensate for the increased leakage that results from a single gate workfunction for both devices.

It should be noted, however, that since the NFET and PFET halo implants are performed after silicidation, these implanted species also end up in the FUSI gates. In the case of the process flow used for this study, phosphorus ends up in the FUSI gate for the NFET, while boron ends up in the FUSI gate for the PFET. At the time of device design and fabrication, it was not known if and to what extent these dopants might shift the gate workfunctions for each device, consequently increasing leakage (e.g., for the NFET, the

gate workfunction would decrease due to the introduction of phosphorus). As both the NFET and PFET gates are fully silicided, though, it was hypothesized that the dopants would distribute evenly throughout the gate material, as ideally there would be no unreacted regions of polysilicon for said dopants to segregate into, and so the gate workfunctions would be largely unaffected. This hypothesis is given more weight when one considers Figs. 5.4 and 5.5 in Chapter 5, whereby the peak ion concentration, relative to the combined atomic concentrations of Ni and Si in NiSi, is on the order of 0.2 %. If the NiSi structure is maintained after the ITS process, then the implanted species would exist at the NiSi grains as "interstitials" and, as they constitute such a small percentage of the gate material, they should not appreciably effect the gate workfunction. One would thus contend that their only contribution to the device on and off states would be the electric field contribution from the implanted ions that diffused toward the NiSi-gate-dielectric interface during the post-ITS thermal step. As these ions are positively charged, such an effect may serve to offset the threshold voltage shift due to interface charge at the body-to-gate-dielectric interface, depending on how many ions diffuse to the bottom of the FUSI gate.

Performing dopant implantation *before* silicidation, however, has been shown to shift the gate workfunction, and the extent of this shift depends on the implanted species and dose [13], [14]. This is because the subsequent silicidation step acts as a "snowplow," pushing the implanted dopants to the poly-oxide interface (where in this case at least some of the dopants become *activated*). As no snowplow effect takes place during a post ITS anneal, though (only dopant segregation occurs), there was little reason to suspect a gate workfunction shift for implantation into FUSI gates. Thus, it would

seem that gate workfunction engineering for bulk switching SFETs would require dopant implantation into the polysilicon gate *before* gate patterning and subsequent metal deposition/silicidation. This can become challenging for ultrathin body regions, however, as for FUSI gates the polysilicon thickness must be scaled with the body thickness, which means that the implantation energy must also be scaled down.

One final point of interest for the bulk switching SFETs fabricated for this study is the V_{out} terminal of the designed inverters and ring oscillators. Although the use of a single metal silicide for full CMOS operation allows the silicide to serve four functions, as mentioned earlier (Fig. 6.4), if careful attention is not paid to the post ITS anneal step, counterdoping of the halo regions adjacent to the V_{out} terminal will increase the series resistance of both the pull-up and pull-down networks, thus reducing drive current and consequently circuit speed (Fig. 6.8). This is due to the fact that the V_{out} terminal is shared by both networks, which means that during the ITS processes, both p-type and n-type dopants are introduced into opposite ends of the terminal. During the anneal step, the n-type dopant species forms a halo region where it should, and likewise for the p-type dopant species; however, the dopant distribution of both species also spreads out *within* the metal silicide. Therefore, if the anneal process is long enough and/or if the gate pitch is short enough, each species can propagate to the other end of the silicide and into the opposing halo region. While the dopant concentration at this opposite end of the V_{out} terminal (e.g., boron concentration at the NFET halo region) may be characterized as the tail end of that dopant distribution, and so the extent of counterdoping may be negligible compared to the halo dopant concentration of interest in said region, with enough time the size of that tail may grow and effect a significant change.

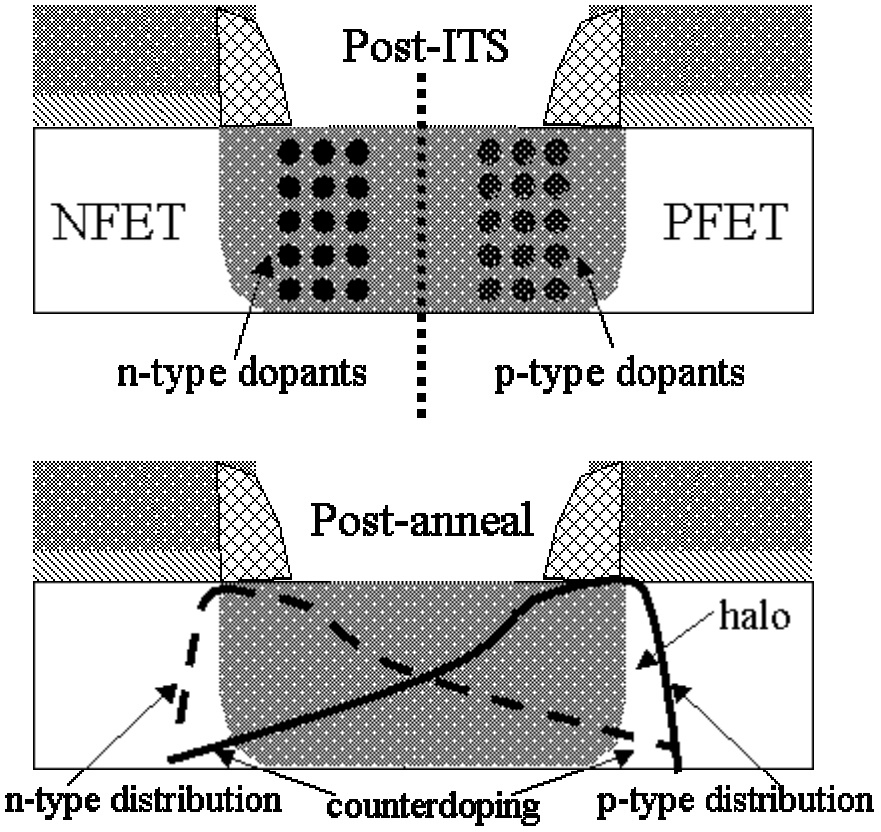

Fig. 6.8. Illustration of counterdoping of the NFET and PFET halo regions at the V_{out} terminal of a single metal Schottky CMOS inverter on SOI substrates after the post-ITS anneal/activation step.

It should be noted that unipolarity is not compromised in the aforementioned instance, though. Of particular concern in the case of the inverter in Fig. 6.4 are the halo regions adjacent to the ground and V_{DD} terminals, and the silicides in said regions are only populated with the dopant species of interest. Thus, a counterdoping of the halo regions adjacent to the V_{out} terminal would not affect the "unipolarity" of the devices within some V_{GS} range – the only change is an increase in series resistance. Consequently, the frequency response is reduced. One may potentially work around this by utilizing SIIS rather than ITS [15], although other challenges such as high-k compatibility remain an open question.

6.3 Test Chip Features

As the primary goal of this study was to empirically demonstrate single metal Schottky CMOS, the test chip design is not particularly complicated. After demonstrating discrete devices, the next step toward a viable CMOS or CMOS replacement technology is to demonstrate the voltage transfer characteristic (VTC) of a simple logic gate, such as an inverter. As such, the majority of the test chip constitutes discrete n-channel and p-channel SFET designs, as well as inverter designs using varying channel widths (1, 5, and 10 μm), NFET:PFET width ratios (1:1, 1:2, and 1:3), and gate lengths (0.5 – 3 μm). Inverters of moderate circuit density (Fig. 6.5) and with somewhat relaxed design rules (Fig. 6.9) were both included in the test chip.

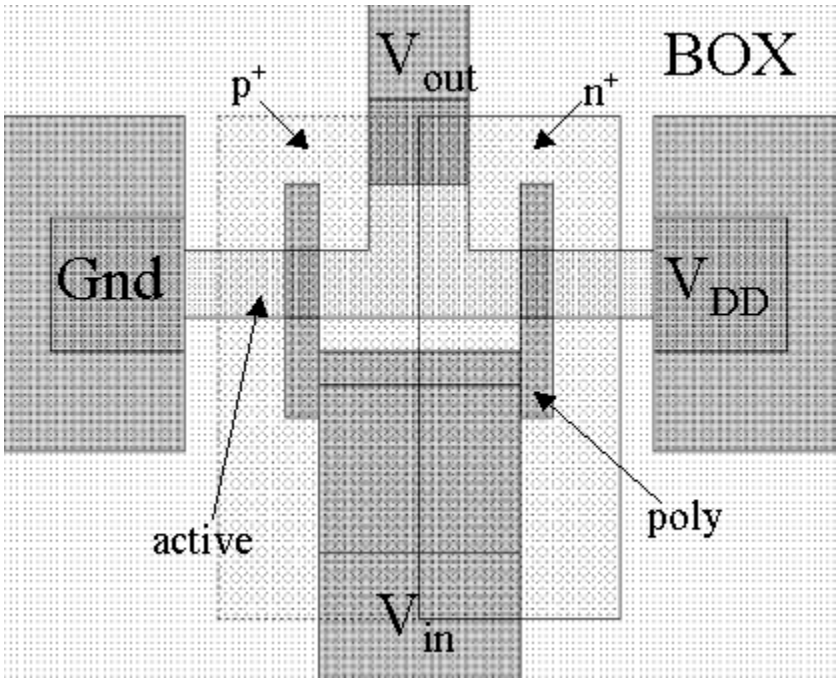

Fig. 6.9. Schottky CMOS inverter layout using relaxed design rules. The p+ and n+ implant windows were enlarged, as was the poly-to-active overlap and the poly-to-M1 spacing.

Beyond inverters, a number of ring oscillators were also designed. These are the most complicated circuits on the test chip, all of which consist of 17 stages. The device length and width ratios were varied much like what was done with the single inverters, but the primary goal here was to demonstrate the achievable circuit density when the silicide source/drain region serves all of the four purposes mentioned previously. An example of this is shown in Fig. 6.10 for inverters with a 1:2 NFET:PFET width ratio. In the case of this type of ring oscillator, the ground and V_{DD} regions are mostly connected at the [silicided] active level (a serpentine pattern ring oscillator would require a metal level to connect every other V_{DD} and ground region, as shown in Fig. 6.10 for the V_{DD} regions), and each inverter stage is isolated from the next by said active level. No specific design rules were defined for the size of the active regions that constitute the ground and V_{DD} regions, so the design in Fig. 6.10 may indeed have increased density if said regions are narrower.

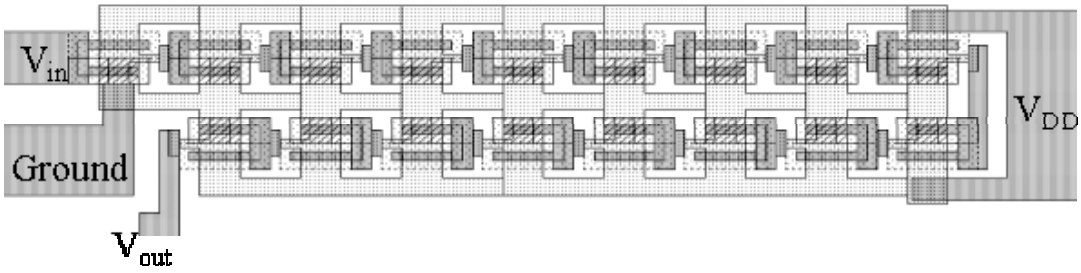

Fig. 6.10. High density 17-stage ring oscillator design using single metal Schottky CMOS. Inter-circuit isolation between each row of inverter stages is achieved with the same silicide that acts as the source/drain regions.

Amongst the usual test structures (i.e., Van der Pauw, Cross-Bridge Kelvin Resistors, SEM structures, alignment verniers, etc.) designed for a test chip, a series of diodes were also included. These diodes are of a fairly simple design – a 100 μm x

100 μm contact probe test pad placed over a slightly larger [silicided] active region – but the purpose for this was to have a "sanity check" for the NiSi-to-Si interface. In other words, it was the intent to have a measurable individual Schottky diode to extract barrier height and diode leakage information from. Diode leakage information is particularly useful in determining whether or not the diode is an actual diode, or whether it is merely an ohmic contact. This is important when analyzing data from the bulk switching SFETs, as a higher than expected drive current (or leakage current) may be attributable to a less than ideal M-S junction (e.g., high diode leakage induced by defects) rather than a larger than expected discrepancy between model results and experimental data.

Chapter 6 References

[1] J.R. Tucker, "Schottky Barrier MOSFETs for Silicon Nanoelectronics," *Proc. IEEE*, 1997, pp. 97-100.

[2] A. Brand, A. Haranahalli, N. Hsieh, Y.C. Lin, G. Sery, N. Stenton, B.J. Woo, S. Ahmed, M. Bohr, S. Thompson, S. Yang, "Intel's 0.25 Micron, 2.0Volts Logic Process Technology," *Intel Technology Journal*, Q3 1998, pp. 1-9.

[3] T. Seki, E. Itoh, C. Furukawa, I. Maeno, T. Ozawa, H. Sano, N. Suzuki, "A 6-ns 1-Mb CMOS SRAM with Latched Sense Amplifier," *IEEE Journal of Solid-State Circuits*, Vol. 28, no. 4, 1993, pp. 478-483.

[4] F. Brady, B. Keshavan, L. Rockett, "Evaluation of the Performance and Reliability of a 1M SRAM on Fully-Depleted SOI," *Proceedings of the 1998 IEEE International SOI Conference*, 1998, pp. 129-130.

[5] H. Pilo, S. Lamphier, F. Towler, R. Hee, "A 300MHz, 3.3V 1Mb SRAM Fabricated in a 0.5μm CMOS Process," *IEEE International Solid State Circuits Conference*, 1996, pp. 148-149, 433.

[6] R.D.J. Verhaar, R.A. Augur, C.N.A. Aussems, L. de Bruin, F.A.M. Op den Buijsch, L.W.M. Dingen, T.C.T. Geuns, W.J.M. Havermans, A.H. Montree, P.A. van der Plas, H.G. Pomp, M. Vertregt, R. de Werdt, N.A.H. Wils, P.H. Woerlee, "A 25 μm^2 Bulk Full CMOS SRAM Cell Technology with Fully Overlapping Contacts," *IEEE IEDM*, 1990, pp. 473-476.

[7] S. Rusu, "Trends and Challenges in VLSI Technology Scaling Towards 100 nm," *Intel Corp.*, 2001. Available: http://www.imec.be/esscirc/esscirc2001/C01_Presentations/404.pdf

[8] F. Deng, K. Ring, Z. F. Guan, S. S. Lau, W. B. Dubbelday, N. Wang, K. K. Fung, "Structural investigation of self-aligned silicidation on separation by implantation oxygen," *J. Appl. Phys.*, Vol. 81, no. 12, 1997, pp. 8040-8046.

[9] F. Deng, R. A. Johnson, P. M. Asbeck, S.S. Lau, W. B. Dubbelday, T. Hsiao, J. Woo, "Salicidation process using NiSi and its device application," *J. Appl. Phys.*, Vol. 81, no. 12, 1997, pp. 8047-8051.

[10] B.-Y. Tsui, C.-P. Lin, "Process and Characteristics of Modified Schottky Barrier (MSB) p-Channel FinFETs," *IEEE Trans. Elec. Dev.*, Vol. 52, no. 11, 2005, pp. 2455-2462.

[11] R.F. Pierret, "Semiconductor Device Fundamentals," *Addison-Wesley Publishing Company, Inc.*, 1996, p80.

[12] B.-Y. Tsui, C.-P. Lin, “A Novel 25-nm Modified Schottky Barrier FinFET With High Performance,” *IEEE Elec. Dev. Lett.*, Vol. 25, no. 6, 2004, pp. 430-432.

[13] J. Kedzierski, D. Boyd, C. Cabral, Jr., P. Ronsheim, S. Zafar, P. M. Kozlowski, J. A. Ott, M. Ieong, “Threshold Voltage Control in NiSi-Gated MOSFETs Through SIIS,” *IEEE Trans. Elec. Dev.*, Vol. 52, no. 1, 2005, pp. 39-46.

[14] W.P. Maszara, Z. Krivokapic, P. King, J.-S. Goo, M.-R. Lin, “Transistors with Dual Work Function Metal Gates by Single Full Silicidation (FUSI) of Polysilicon Gates,” *IEEE IEDM*, 2002, pp. 367-370.

[15] A. Kinoshita, C. Tanaka, K. Uchida, J. Koga, “High-performance 50 nm-gate-length Schottky-source/drain MOSFETs with dopant segregation technique,” *VLSI Symp. Tech. Dig.*, 2005, pp. 158-159.

Chapter 7

Silicon-on-Insulator (SOI) SFETs and CMOS Implementation

7.1 Method of Fabrication

The process flow started with p-type (boron-doped, 14-22 Ω-cm, 4-8x10^{14} cm^{-3}) UNIBOND™ Smart Cut™ SOI wafers with a body thickness, t_{body}, of 100 nm and a buried oxide (BOX) thickness of 200 nm. After defining the active regions, phosphorus was implanted into selected regions for n-well formation at 50 keV with a 5x10^{11} cm^{-2} dose and a subsequent 4 h furnace anneal at 1000 °C in N_2. Simulation in Silvaco Athena showed the resultant n-well profile to be uniform throughout the entire body region with a concentration of 1x10^{15} cm^{-3} after all thermal processing. A 9 nm gate oxide was thermally grown, after which 130 nm of undoped polysilicon with a 100 nm nitride cap were deposited via LPCVD. After gate patterning, a 30 nm thick oxide sidewall spacer was grown. The oxide over the source/drain regions was then removed in a dry etch with CHF_3 and O_2 and the nitride cap was stripped in phosphoric acid at 175 °C.

A 30 s, 50:1 HF dip, followed by a 1 min rinse in DI water and then a spin rinse/dry, was performed. The wafers were immediately loaded into a sputter chamber and placed under vacuum. Nickel was then sputter deposited to ~ 45 nm after reaching a base pressure of 1-2 μTorr. The silicidation step was performed at 500 °C for 1 min in N_2 via RTA, and unreacted nickel was removed in a 2:1 H_2O_2:H_2SO_4 mixture at 90 °C. Although the two-step silicidation process in [1]-[5] is more suitable for aggressive scales, in this study, the devices are large enough to warrant a one-step silicidation without the risk of shorting across the body region. Another purpose of the one-step

silicidation was to achieve a direct gate overlap to the M-S junctions, as such a one-step process has been shown to result in greater lateral silicidation than a two-step process [2]. This increases current injection at the M-S junctions, as the gate field is superimposed with the halo-induced field.

Split	Fluorine co-implant (y/n)	Post-ITS anneal (°C/min)	N/PFET well type	N/PFET W_{halo} (nm)	N/PFET *EOT* (nm)
1	N	600/30	P/N	21.6/19	18.4/19.2
2	Y	600/30	N/P	19.4/16.4	17.6/18.6
3	Y	700/30	P/N	31/17.5	15/19.8

Table 7.1. ITS splits, well type, *EOT*, and W_{halo} results.

An ITS process was performed for both the NFETs (phosphorus implant) and the PFETs (BF_2 implant). For all implants, the dose and energy were $4x10^{15}$ cm^{-2} and 34 keV, respectively. To form the halo regions, a subsequent thermal anneal was performed via RTA at 600 °C or 700 °C for 30 min in 10 min pulses (Table 7.1). For some splits, fluorine was blanket implanted ($4x10^{15}$ cm^{-3} @ 34 keV) before the n- and p-type halo implant windows were defined. The primary purpose of the fluorine implant was to increase the thermal stability of the NiSi (from 600 °C to 750 °C) [6], thus reducing defects at the M-S junction and potentially facilitating higher active dopant concentrations in the halo regions (i.e., higher drive current). While fluorine is already present in the BF_2 implant to serve this purpose, it is not present during the phosphorus implant, and so the upper thermal limit is restricted by the NFETs to 600 °C due to silicide agglomeration [6]. The size of the halo regions was approximated electrically using capacitance-voltage (C-V) analysis and some simple assumptions (shown in the next section).

After the halo formation, aluminum metallization was performed with an evaporation/liftoff process using Clariant nLOF 2020 resist and AZ-300T resist stripper. A top-down picture of the final circuit structure (inverter) is shown in Fig. 7.1, which shows the NFET and PFET sharing a metallic source/drain (MSD) region at the V_{out} terminal, as discussed in Chapter 6.

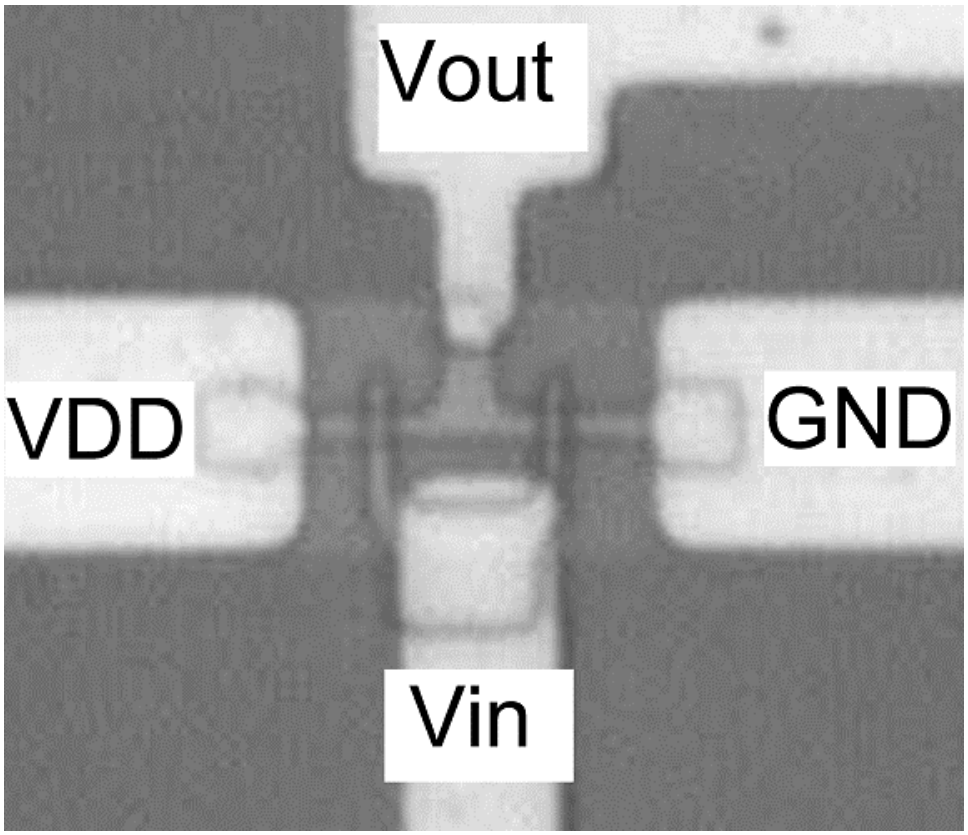

Fig. 7.1. Top-down picture of MSD CMOS inverter after all fabrication was completed. The mask-defined gate length and width are 0.5 μm and 1 μm, respectively.

7.2 Extracting the Halo Width (W_{halo})

In the modeling study in Chapter 5, it was suggested that dopants implanted into the silicide redistribute throughout the entire silicide and segregate to the M-S interface very quickly during the post-ITS anneal. This information may prove useful for a device

structure where the extent of lateral and vertical diffusion for the as-implanted dopants to redistribute throughout the silicide is about the same, as it allows one to use a test structure reliant on purely vertical diffusion to estimate the width of the lateral halo region extending from the source/drain silicide. Such a measurement might be performed through SIMS (secondary ion mass spectroscopy) or SRP (spreading resistance profiling) analysis, as they give information on the size of the halo region, the exact halo profile, and the dopant concentration at the M-S interface. Without access to such measurement tools, however, one is forced to be more creative in their characterization techniques. This is where capacitance-voltage (C-V) measurements become very useful.

For the process flow discussed in Section 7.1, FUSI gates were not formed, due to the gate polysilicon being too thick for the deposited nickel thickness to fully consume. This is reflected in the *EOT* results in Table 7.1, which are ~ 2x larger than the physical gate oxide thickness of 9 nm. Although this reduces device performance compared to an ideal situation (i.e., FUSI gates), it does allow one to electrically extract W_{halo} using C-V analysis. Knowing the as-deposited nickel and polysilicon thicknesses, and considering the 1:1.84 Ni:Si consumption ratio to form NiSi [7], one can calculate the thickness of the unconsumed polysilicon which acts as a capacitor in series with the gate oxide capacitance. After finding the capacitance in the accumulation region of the C-V curve, one can electrically extract W_{halo} using:

$$W_{halo} = d_u - \varepsilon_0 \varepsilon_{si} A * \left(\frac{1}{C_{acc}} - \frac{1}{C_d} \right) \quad (1)$$

where d_u is the thickness of the unconsumed polysilicon, C_{acc} is the accumulation mode capacitance, C_d is the gate dielectric capacitance, ε_0 is the vacuum permittivity, ε_{si} is the relative silicon permittivity, and A is the capacitor area.

W_{halo} results for all three splits are shown in Table 7.1. It is noted that one or two of the cleanest samples were used for each data point, although realistically there should exist some variability associated with this measurement. The key assumptions of this C-V method are: 1.) diffusion within silicon and polysilicon are the same in the temperature range of interest; 2.) uniform silicide diffusion front (i.e., silicide spiking is very small compared to the unconsumed polysilicon thickness); 3.) the purely vertical diffusion within the gate silicide is representative of both the vertical and lateral diffusion within the source/drain silicide; 4.) the silicide in the polysilicon gate is the same phase as that of the source/drain region; 5.) uniform and degenerately doped halo region (i.e., no halo depletion). Assumptions 3-5 have the most significance, as this method measures the width of the *undepleted* portion of the halo region formed by purely vertical diffusion in NiSi formed on a polysilicon gate. It is not unreasonable to suggest that the tail of the halo profile depletes during this measurement, resulting in an underestimation of the actual W_{halo}. Assuming a 1 nm/dec junction abruptness, for a degenerately doped halo region, the tail should extend ~ 4-5 nm before the halo concentration reduces to the 10^{15} cm^{-3} level. In the case of the p-type halo region for the 600 °C/30 min split without the fluorine co-implant (split 1), the actual W_{halo} is therefore ~ 23-24 nm, which is in very good agreement with SRP results [6].

It is interesting that, from Table 7.1, the fluorine co-implant seems to retard both boron and phosphorus diffusion during the post-ITS anneal at 600 °C (split 2 vs. split 1). As discussed in Chapters 3 and 5, dopant diffusion within silicides has been attributed to grain boundary diffusion, and this may explain the observed result for two reasons. First, the fluorine co-implant was performed before the halo implants, and so the damage from

the fluorine implant reduces the effective grain size of the silicide. Although this increases the grain boundary density (presumably increasing diffusion), it also makes the path of the implanted dopants to the M-S interface less direct due to increased scattering as the dopants diffuse from one [smaller] grain boundary to the next. Additionally, the implanted fluorine ions fill up some of these interstitial sites, which the implanted dopants must now compete for. While this feature size reduction may seem beneficial for fabrication at aggressive scales, the co-implanted fluorine itself also seems to reduce PFET performance (shown later). Even for the 700 °C/30 min split with fluorine (split 3), the p-type halo is still smaller than the 600 °C/30min split without fluorine (split 1), while the n-type halo grows considerably in size (in [6], a 700 °C/30 min anneal to BF_2 ITS *without* a fluorine co-implant resulted in a 28 nm W_{halo}). There are a number of possible mechanisms for this, although to a first order, it would seem to indicate a higher interstitial/grain boundary dependence on diffusion for boron than for phosphorus, both within the silicide and at the M-S junction. Another possibility is that the fluorine may be counterdoping some of the p-type halo region, and that this counterdoping has a greater temperature dependence than the growth of the p-type W_{halo}.

7.3 Demonstration of Metallic Source/Drain (MSD) CMOS

Fig. 7.2 shows voltage transfer characteristics (VTCs) for what is, to the best of the author's knowledge, the first-ever full empirical CMOS demonstration with metallic source/drain devices on SOI substrates. These results are from split 1 in Table 7.1, for mask-defined gate lengths, $L_{g,m,}$ from 2 μm down to 0.6 μm, a mask-defined width, W_m,

of 1 μm, and a power supply voltage (V_{DD}) of 3 V. The J_{DS} vs. V_{DS} plots for each device in the 0.6 μm inverter are shown in Fig. 7.3.

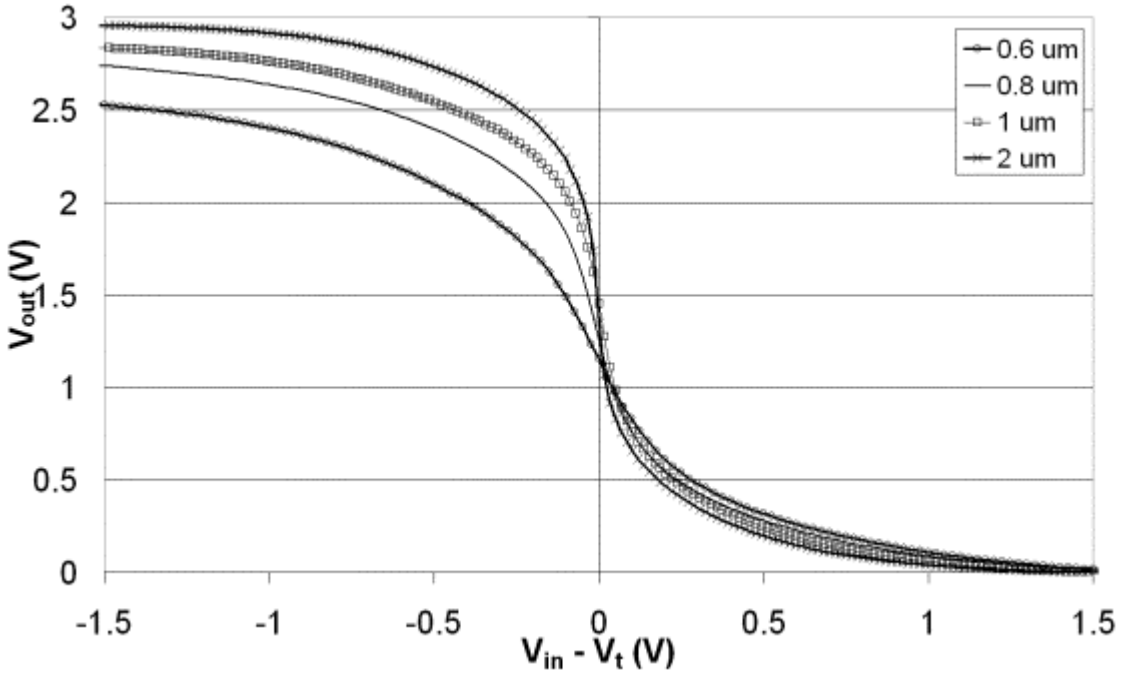

Fig. 7.2. Inverter VTCs for $L_{g,m}$ = 2 μm down to 0.6 μm. The PFET:NFET width ratio is 1:1, and W_m is 1 μm.

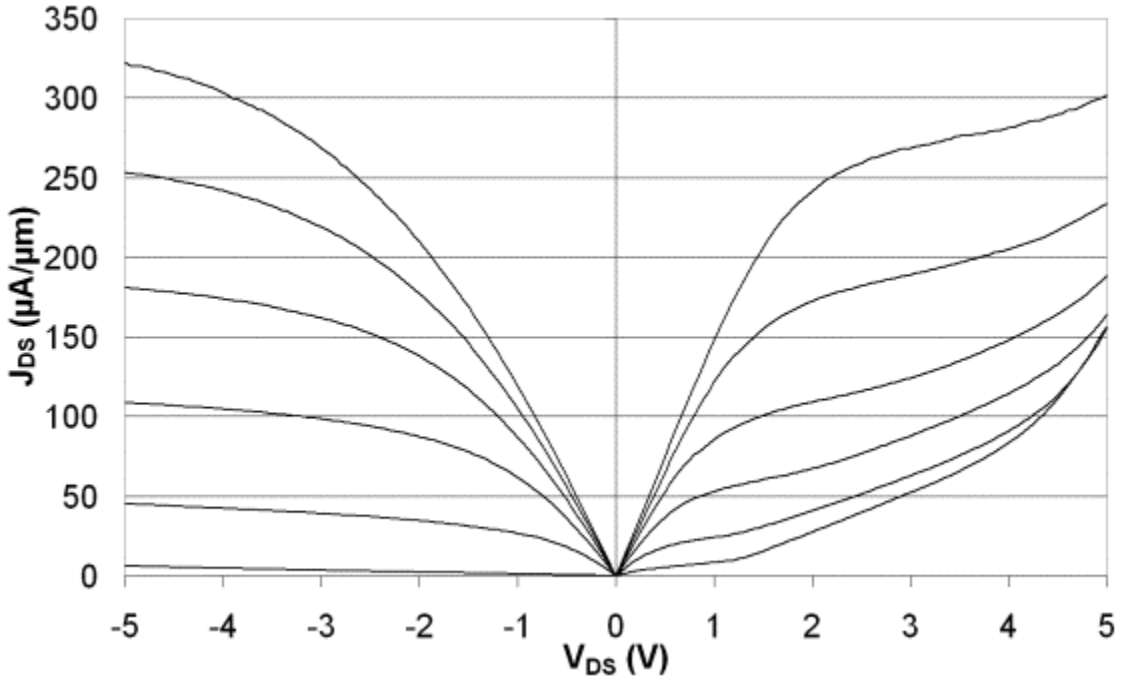

Fig. 7.3. NFET and PFET J_{DS} vs. V_{DS} from $L_{g,m}$ = 0.6 μm inverter in Fig. 7.2. $|V_{GS} - V_{tlin}|$ = 0-5 V in 1V increments.

The poorer pull-up performance relative to the pull-down performance in Fig. 7.2 is due to excessive NFET leakage (Fig. 7.3), which shows punchthrough-like characteristics. For $V_g - V_{tlin} = 0$ V and 1 V, there are two "kinks" in the curve – one at $V_{DS} \sim 1.5$ V and another at $V_{DS} \sim 4.5$ V (punchthrough-like). It is the former kink that seems to be the cause of such anomalous levels of NFET leakage, and this will be explored in more detail in Section 7.4. Suffice it to say that this NFET leakage causes a substantial reduction in inverter gain as $L_{g,m}$ is scaled down (Fig. 7.4). Considering the pull down performance in Fig. 7.2 for all four cases shown, though, it is not unreasonable to suspect much higher inverter gain with an NFET that does not exhibit so much leakage, and that the reduction in this gain with $L_{g,m}$ would not be as severe as what is shown in Fig. 7.4. This is also reflected in Figs. 7.5 and 7.6, where the noise margin low (NML, Fig. 7.5) changes very little if at all with $L_{g,m}$ and the noise margin high (NMH, Fig. 7.6) shows a very noticeable shift at higher V_{DD} values as $L_{g,m}$ drops below 1 μm.

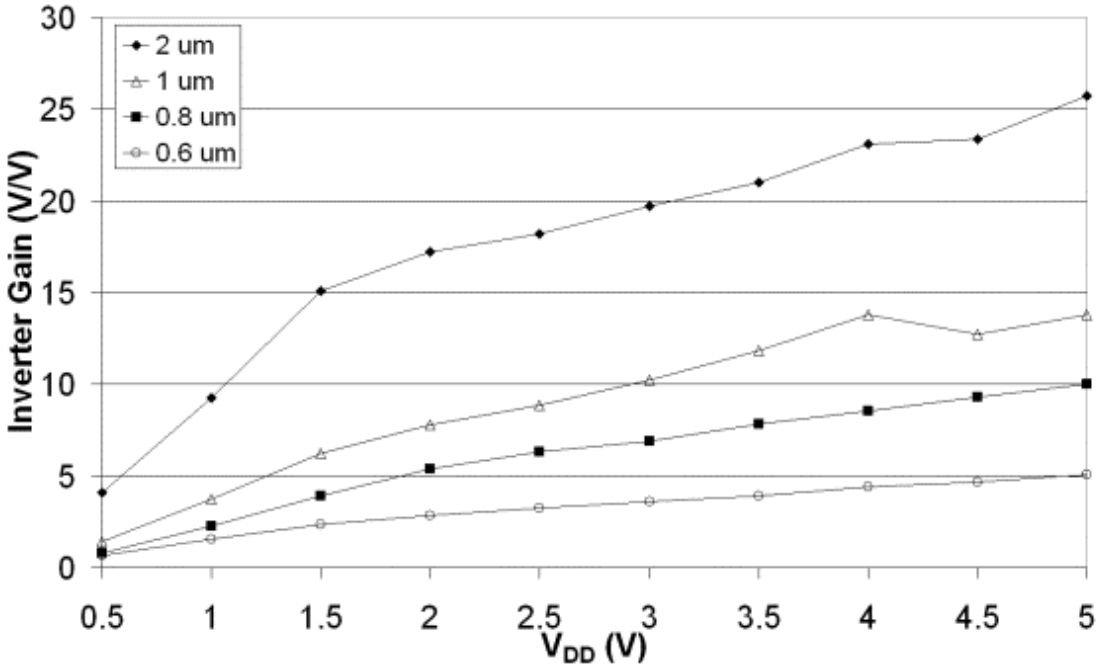

Fig. 7.4. Inverter gain vs. V_{DD} for the inverters from Fig. 7.2.

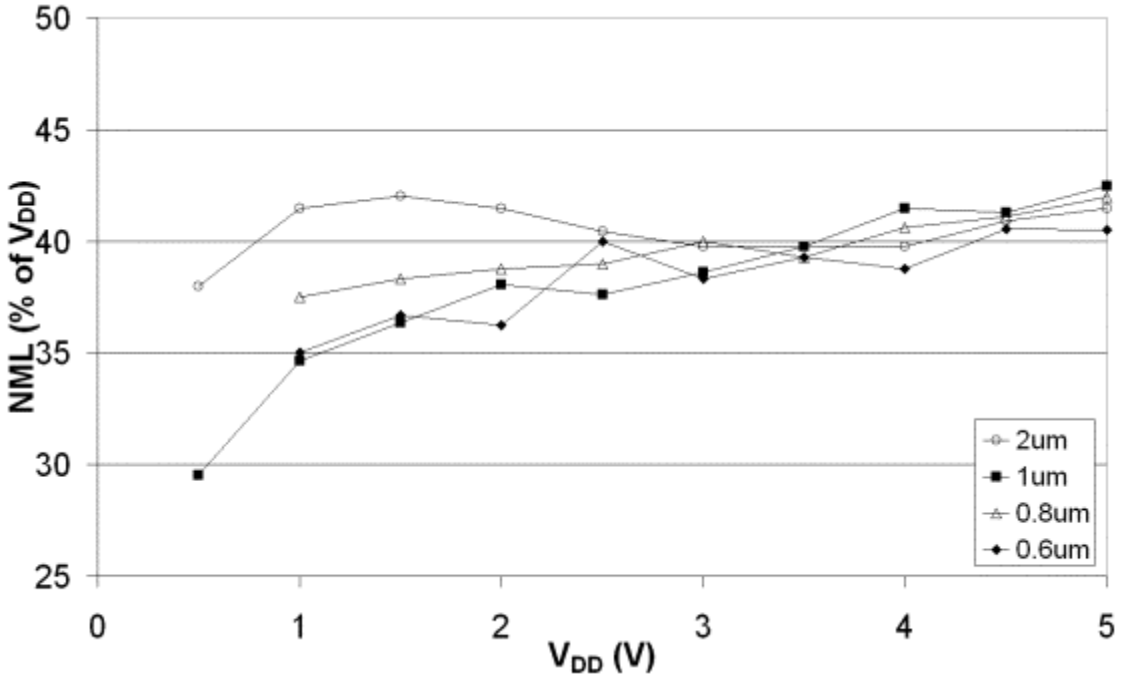

Fig. 7.5. Noise margin low (NML) vs. V_{DD} for the inverters from Fig. 7.2. At moderate to high V_{DD} values, NML has little if any dependence on $L_{g,m}$.

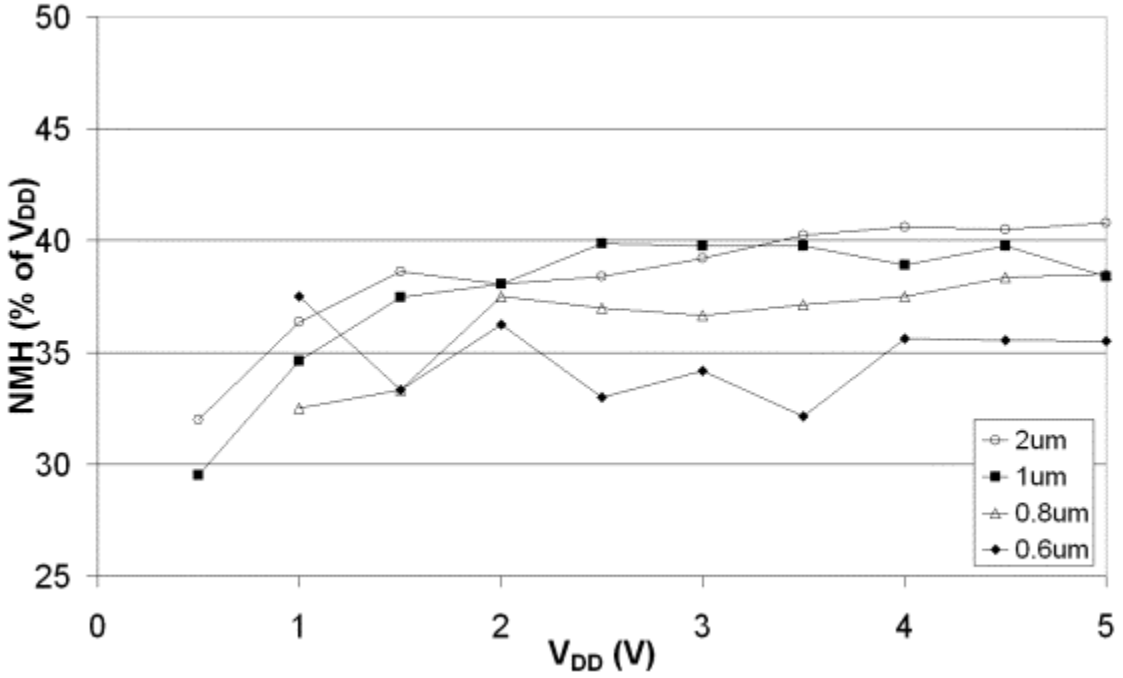

Fig. 7.6. Noise margin high (NMH) vs. V_{DD} for the inverters from Fig. 7.2. As $L_{g,m}$ is scaled below 1 μm, the increasing NFET leakage reduces the range over which a given input is read as a logic high (NMH).

It is noted that Figs. 7.2 – 7.6 are only data from split 1 in Table 7.1. VTCs could not be demonstrated with split 3, because the NFET leakage was actually similar to or greater than the PFET drive current (shown later). As for split 2, most of the samples tested did not perform nearly as well as split 1. An example of this is shown in Fig. 7.7, which compares $L_{g,m}$ = 1 μm inverters from splits 1 and 2 with V_{DD} = 3 V. As can be seen, the gain is substantially lower for the split 2 inverter, and both the pull-up and pull-down operations are farther from V_{DD} and ground, respectively, over a 3 V range of V_{in}. This is attributable to inferior subthreshold swing due somehow to the fluorine co-implant (shown later). Interestingly, one [anomalous] sample from split 2 did indeed yield reasonable NFET performance, whereby the NFET off state current density almost exactly matches that achieved in [4]. This is also shown in Fig. 7.7 in the curve representing the $L_{g,m}$ = 2 μm inverter. The transfer characteristics from this sample are shown in Fig. 7.8, while the J_{DS} vs. V_{DS} is shown in Fig. 7.9. It would seem that the contribution of the subsurface leakage is reduced in this particular NFET. However, that the NFET off state remains flat at ~ 60 pA/μm (for V_{DS} = 0.1 V) while the PFET off state drops below 1 pA/μm suggests that the parasitic leakage mechanism in the NFET has not been completely removed, and that the gate field that would normally effect this current is screened by the surface source/drain silicide due to a recessed silicidation front in the subsurface region. Another interesting point from Fig. 7.8 is the emergence of a dual subthreshold swing for the PFET. This is further evidence of the co-implanted fluorine affecting the p-type halo region by counterdoping and/or reducing diffusion, consequently increasing the influence of the Schottky barrier on current injection. The GIDL-like leakage in the saturation mode curves is due to band-to-band tunneling (BBT)

at the source and drain halo-body barriers, as well as tunneling through the halo regions, and may be a limiting factor to inverter performance at aggressive scales.

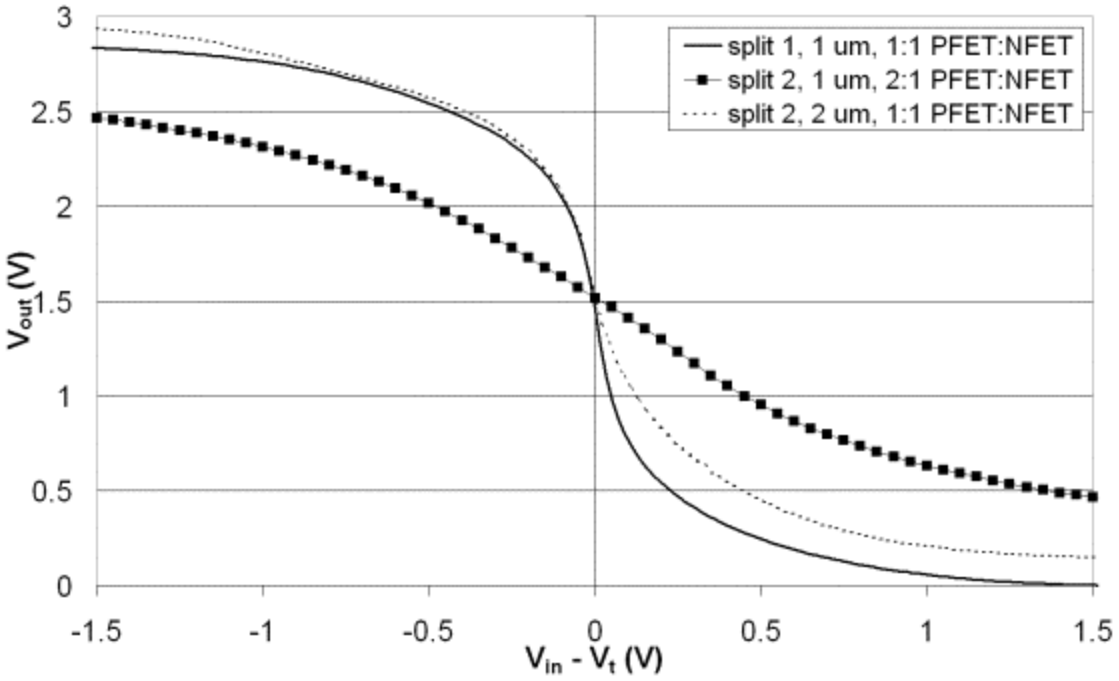

Fig. 7.7. Inverter VTCs comparing splits 1 and 2. The $L_{g,m}$ = 1 μm inverter VTC from split 2 is representative of most of the tested samples in split 2.

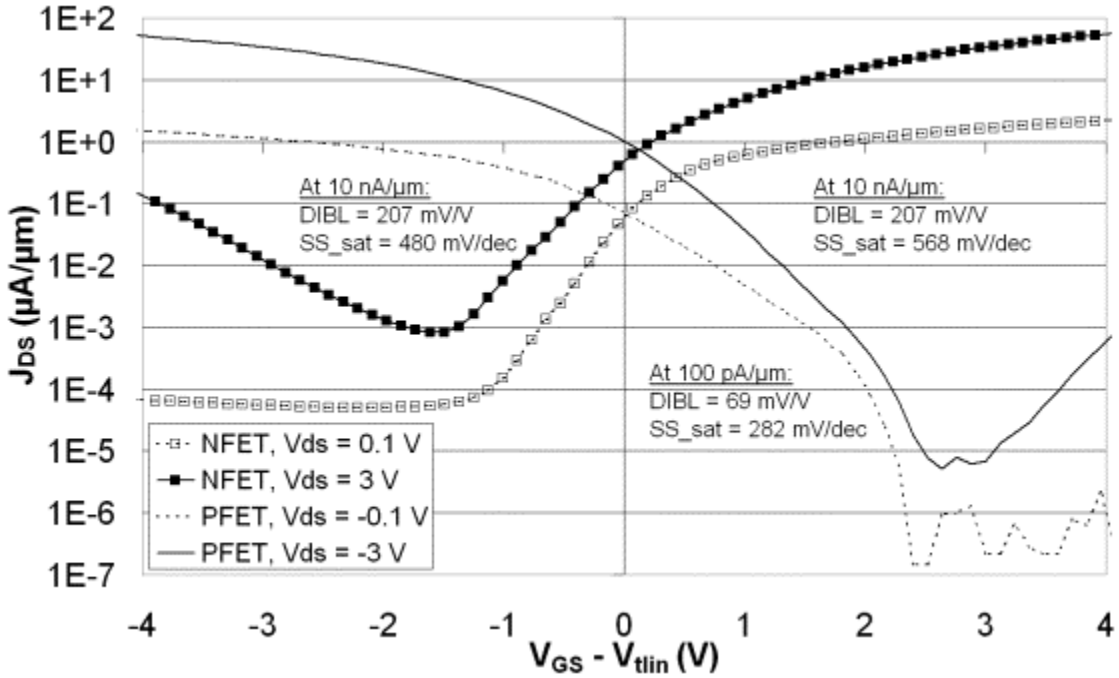

Fig. 7.8. NFET and PFET transfer characteristics from anomalous sample in split 2.

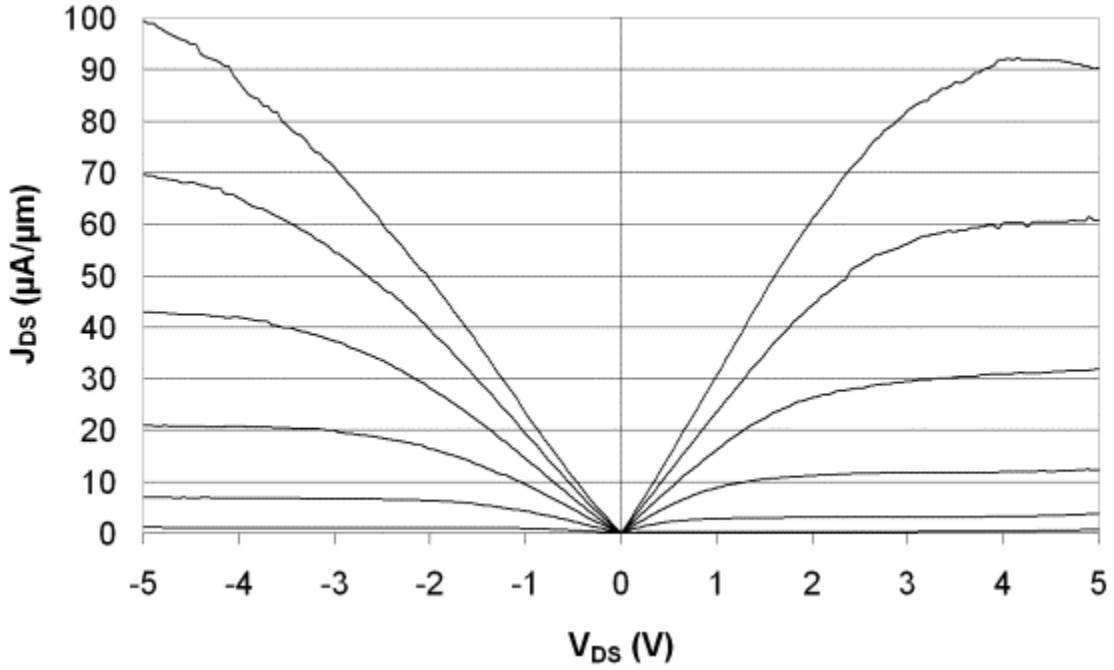

Fig. 7.9. NFET and PFET J_{DS} vs. V_{DS} from anomalous sample in split 2. $|V_{GS} - V_{tlin}|$ = 0-5 V in 1 V increments.

7.4 Analysis of NFET Leakage

The observed NFET leakage in most of the samples tested is by no means a limitation of the basic device structure, as acceptable NFET performance was achieved in [4] using a similar process flow and with much smaller devices. However, said study used a thinner body region (40 nm) and a similar ITS energy (30 keV), which suggests that the NFET performance may be a function of the as-implanted phosphorus profile and whether this profile extends the full depth of the silicide. TRIM simulation (Fig. 7.10) shows that, at both 30 and 34 keV, most of the implanted phosphorus ions are contained within the first 60 nm of the silicide film, which leaves ~ 40 nm of unoccupied NiSi in the subsurface region of the NFETs before the post-ITS anneal. For the ITS process in [4], the as-implanted dopants were spread throughout the entire silicide (Fig. 7.10), and

while the NFET performance is subsequently improved compared to what is presented here, it still did not match up to the PFET performance in that study.

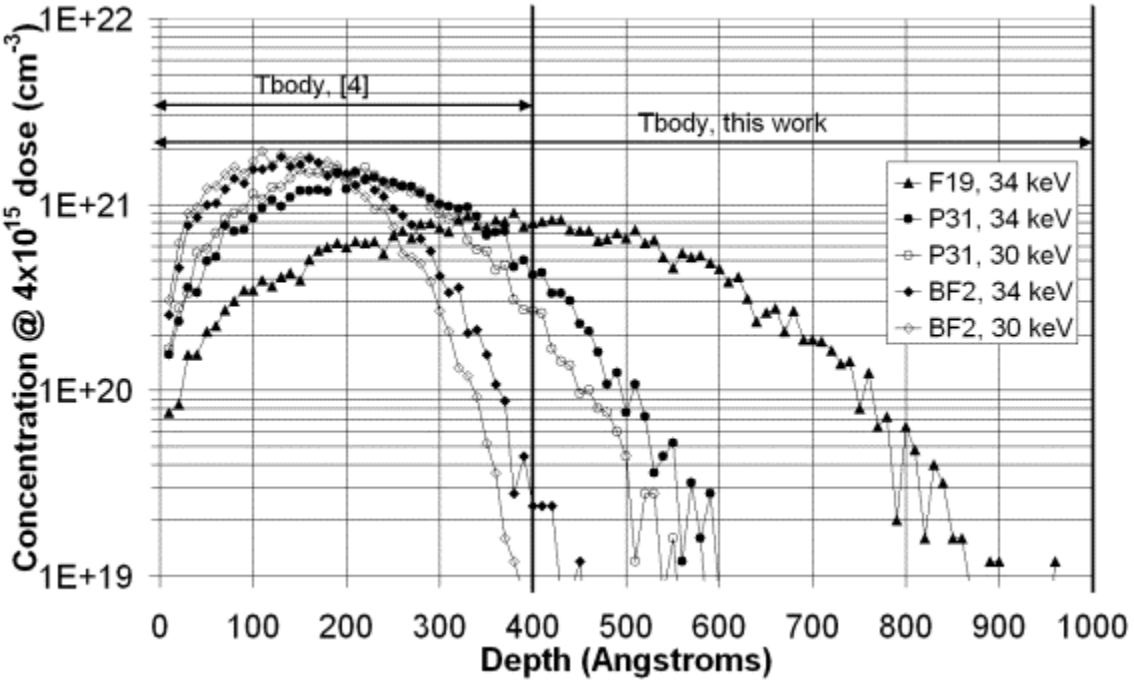

Fig. 7.10. TRIM results for phosphorus, BF_2, and fluorine implants into NiSi at 30 and 34 keV with a $4x10^{15}$ cm^{-2} dose and a 10,000 ion count.

Some more insight into the physical mechanisms of the NFET leakage is gained through C-V analysis. The C-V structure used has a 500 μm x 500 μm gate, surrounded by the source/drain silicide. The halo region between this silicide and the body region acts as the body contact. Fig. 7.11 shows normalized C-V curves for structures with the n-type halo contacting the n-type body region and the p-type halo contacting the p-type body region after a post-ITS anneal at 600 °C for 30 min (no fluorine co-implant). The kinks in the C-V curves at V-V_{mid} ~ 1-2 V at 100 kHz for the p-type and n-type structures are attributable to donor-like and acceptor-like states at the oxide-silicon interface, respectively. More importantly, though, is the behavior of the C-V curves in depletion mode. While the p-type structure depletes fully, the n-type structure does not. The

TRIM results in Fig. 7.10 rule out the possibility of the phosphorus implant punching through the gate stack (thus increasing the n-type body doping and therefore NFET leakage), which is further substantiated by a measured *EOT* that is ~ 2x larger than the physical oxide thickness. This singles out the source/drain capacitance as the culprit. The frequency dependence of these curves suggests that the NFET source/drain capacitance is artificially high due to defect-induced leakage at the M-S junction toward the source/drain-BOX interface, perhaps due to silicide agglomeration in this region. This is also supported by Fig. 7.3, where the NFET saturation region exhibits some curvature. If such leakage were due to SCE, this saturation region would be sloped, but without curvature (i.e., constant slope). That the slope is changing in Fig. 7.3 indicates a parasitic diode shunting the halo region formed at the surface. Indeed, before the first kink at V_{DS} ~ 1.5 V in Fig. 7.3, the NFET actually behaves rather normally.

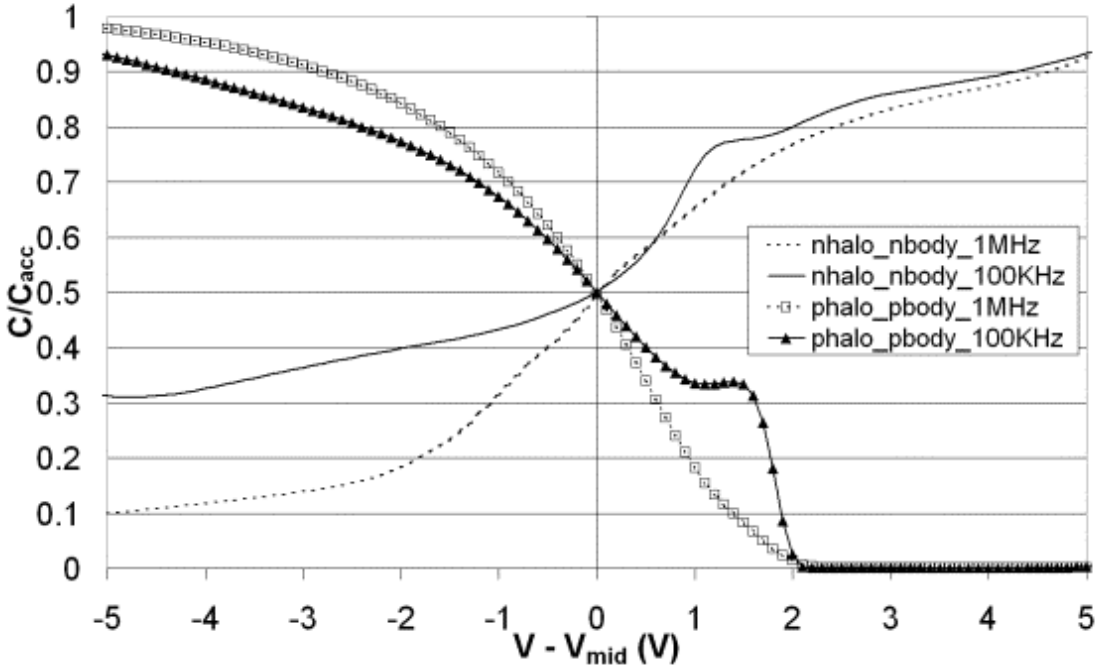

Fig. 7.11. Normalized C-V curves for n-type and p-type structures at 100 kHz and 1MHz. C_{acc} is the accumulation mode capacitance and V_{mid} is where C/C_{acc} = 0.5.

Fig. 7.12 shows C-V curves at 100 kHz for four process splits with n-type halo regions contacting the body region. The higher depletion mode capacitance for both p-body splits, compared to the n-body splits, may be attributed to higher hole leakage through and over the lower hole SBH than for electrons at a NiSi-Si Schottky junction, whereby the unmodified electron SBH is 0.65 eV [8] and is ~ 0.47 eV for holes. This suggests that an n-body accumulation mode NFET may actually outperform a p-body inversion mode NFET for this device structure, due to the higher subsurface barrier height to majority carriers in the body region.

It is very interesting that the PFET exhibits better performance than the NFET (Fig. 7.3), even with a smaller projected range into the NiSi for BF_2 compared to phosphorus (Fig. 7.10). This is at least partly attributable to the fluorine from the BF_2 implant, as suggested in Section 7.2. This is also reflected in Fig. 7.12, where both the n-body and p-body structures (both with n-type halo regions) with the fluorine co-implant do not exhibit the deep depletion-like characteristic that the splits without the fluorine co-implant show. This indicates that the fluorine reduces the defect concentration at and near the M-S junction, although the depletion mode capacitance is still not reduced. That the fluorine has a stronger effect on the n-well split than the p-well split may be due to a difference in the concentration of donor-like and acceptor-like states at the M-S interface. As these states fill up with electrons or holes, the effective barrier height to the respective carrier is increased, thus decreasing the effective capacitance, as Fig. 7.12 shows for the splits without the fluorine co-implant. This all suggests that the n-type halo is not uniformly distributed throughout the body thickness. This would cause the subsurface source/drain capacitance to be dominated by a higher Schottky capacitance that shunts

the lower source/drain capacitance at the surface, which is dominated by the halo-body capacitance. Assuming this is the case and using the equation for Schottky diode capacitance at zero bias [9], the effective dopant concentration in the subsurface region is ~ 2.5×10^{16} cm^{-3} (n-type) for the n-body structure and ~ 1.4×10^{16} cm^{-3} (p-type) for the p-body structure. These are actually overestimates, as the halo region lowers the SBH, which increases the capacitance achieved for a given dopant concentration. Therefore, it is highly likely that the n-type halo implant does not even propagate toward the M-S interface at the bottom of the silicide during the post-ITS anneal. NiSi phase has been shown to be a function of depth, whereby the silicon-rich phase exists toward the bottom of the silicide [10], as well as whether the silicidation takes place in silicon or polysilicon due to stress effects [2]. Thus, it is not unreasonable to suggest lower phosphorus diffusion than boron in the subsurface source/drain silicide. This is where the assumption of silicide phase similarity between silicon and polysilicon may fall apart.

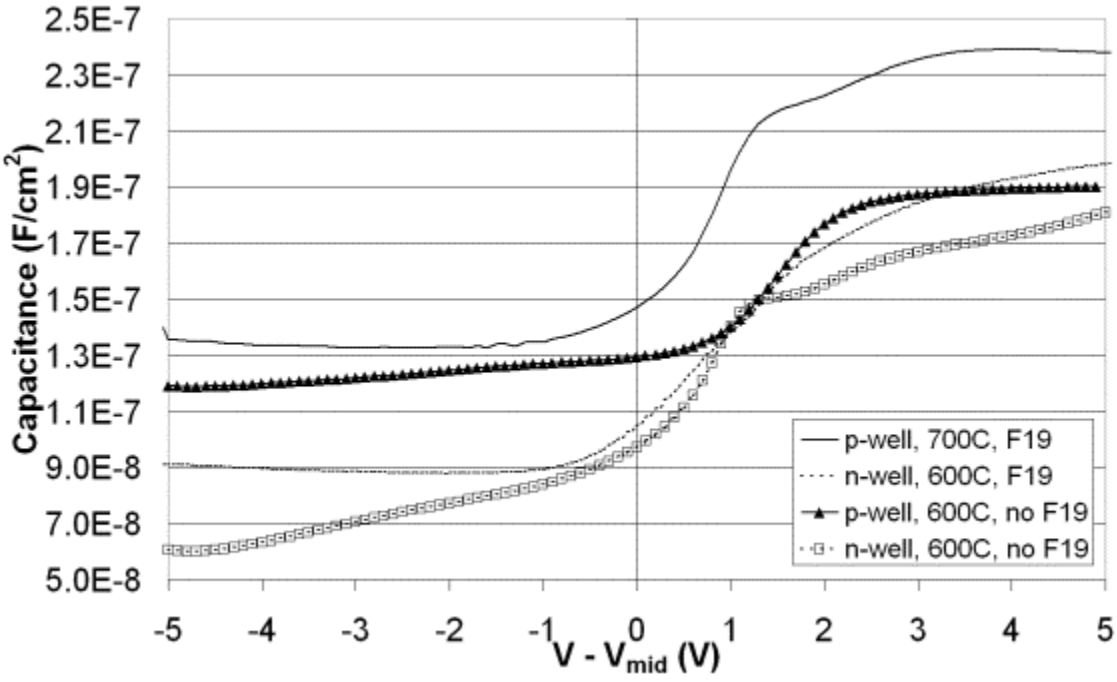

Fig. 7.12. Normalized C-V curves for various n-type halo structures at 100 kHz.

This analysis of NFET leakage lends further weight to the notion in [4] that phosphorus diffusion (or lack thereof) within NiSi is a limiting factor to NFET performance in this structure. That the linear regions for the NFET and PFET in Figs. 7.3 and 7.9, where the n-type and p-type ITS doses were equal, are similar further support the claim in [4] that the phosphorus dose plays a role in current injection through and over the Schottky barrier. This role is equally important as the role that the FUSI source/drain thickness (i.e., body thickness) plays in NFET leakage. At this point, it is reasonable to suspect that the formation of the n-type halo region has a significant contribution from the lateral projected range and straggle from the ITS process, as this results in some portion of the lateral implant tail being injected into the silicon at the M-S interface.

It would seem that improving NFET performance in future efforts with the presented device and circuit structure would require a thinner FUSI source/drain region and/or a higher implant energy for the n-type halo region. It is noted, though, that a higher implant energy will spread out the as-implanted profile, potentially reducing the halo concentration at the M-S interface for a given dose. This is avoided with a thinner body region, as lower implant energies can be utilized to increase the halo concentration throughout the entire halo depth by effectively "squeezing" the as-implanted profile. Another approach would be to utilize SIIS (silicidation-induced impurity segregation) rather than ITS, presumably avoiding the subsurface leakage issue altogether if dopant segregation at the silicidation front is high enough. SIIS would also permit the use of heavier n-type dopants, such as arsenic and antimony, that would not perform as well for an ITS process due to their bigger size, but have been shown to result in superior workfunction modification during SIIS processing [11].

7.5 Effect of Fluorine Co-Implant on Device Performance

Figs. 7.13 and 7.14 show the NFET J_{DS} vs. V_{DS} and transfer characteristics, respectively, for the three process splits in Table 7.1, while Figs. 7.15 and 7.16 show the same for the PFET. For both devices, $L_{g,m}$ = 2 μm and W_m = 1 μm. Split 3 is not shown in Fig. 7.14, as it is little more than a log-linear curve, which Fig. 7.13 already suggests. As is shown in Fig. 7.12, the fluorine co-implant is not expected to reduce the NFET off state current by much if at all in this structure, which is substantiated by Fig. 7.14. Figs. 7.13 – 7.16 suggest that both the NFET and PFET achieve higher drive current without a fluorine co-implant for a given post-ITS anneal. Again, this is attributable to the reduced dopant diffusion within the silicide caused by the implanted fluorine. For the NFET, the fluorine co-implant actually increases DIBL and SS, while only SS is increased for the PFET. This is explained by fluorine acting as an n-type dopant in addition to a diffusion inhibitor, spreading out the tail of the n-type halo region (while also lowering the interface concentration) and counterdoping and/or reducing diffusion of the p-type halo region. As a result, for the PFET, the influence of the Schottky barrier at the M-S junction is increased, consequently increasing SS (Fig. 7.16). For the NFET, the influence of the Schottky barrier is also increased, as shown in Fig. 7.13, where the curves for split 2 exhibit some sub-linear behavior.

The original purpose of the fluorine co-implant, however, was to facilitate higher temperature post-ITS anneals, and the higher temperature may outweigh any adverse effect of the fluorine on device performance. At least in terms of drive current, this seems to be the case for the NFET (Fig. 7.13). Split 3 restores the performance in the linear region compared to split 2, even improving on what is achieved with split 1 with

much higher drive current. However, the device does not reach saturation, again due to subsurface leakage through and over the subsurface Schottky barrier.

The exact opposite effect of fluorine happens for the PFET at 700 °C, however (Figs. 7.15 and 7.16). While the DIBL and SS performance of split 3 is comparable to split 2, the drive current is considerably lower than both splits 1 and 2. That the drive current changes inversely with temperature suggests a stronger thermal activation dependence for fluorine than boron in the temperature range of interest and that, again, the fluorine is counterdoping the p-type halo region. As a result, inverter VTCs could not be demonstrated with split 3, as the PFET drive capability cannot outweigh the NFET leakage. A well-engineered metallic source/drain (MSD) CMOS circuit, then, would mask the fluorine co-implant from the PFETs.

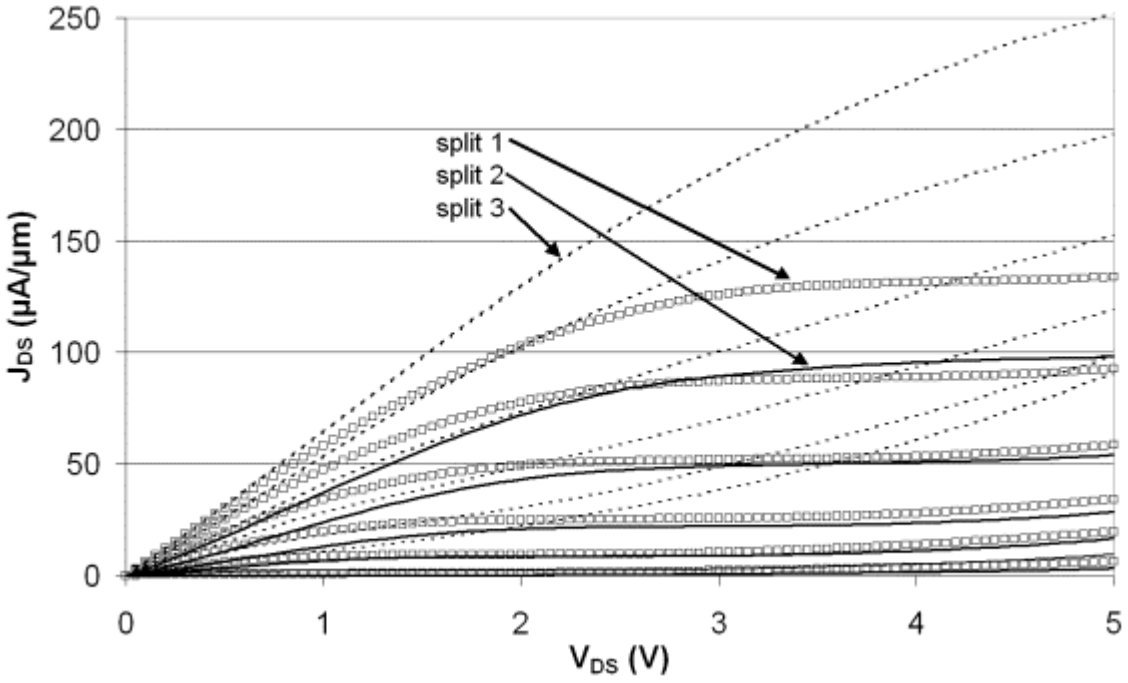

Fig. 7.13. NFET J_{DS} vs. V_{DS} for splits 1-3. $L_{g,m}$ = 2 μm, W_m = 1 μm, and $V_{GS} - V_{tlin}$ = 0-5 V in 1 V increments.

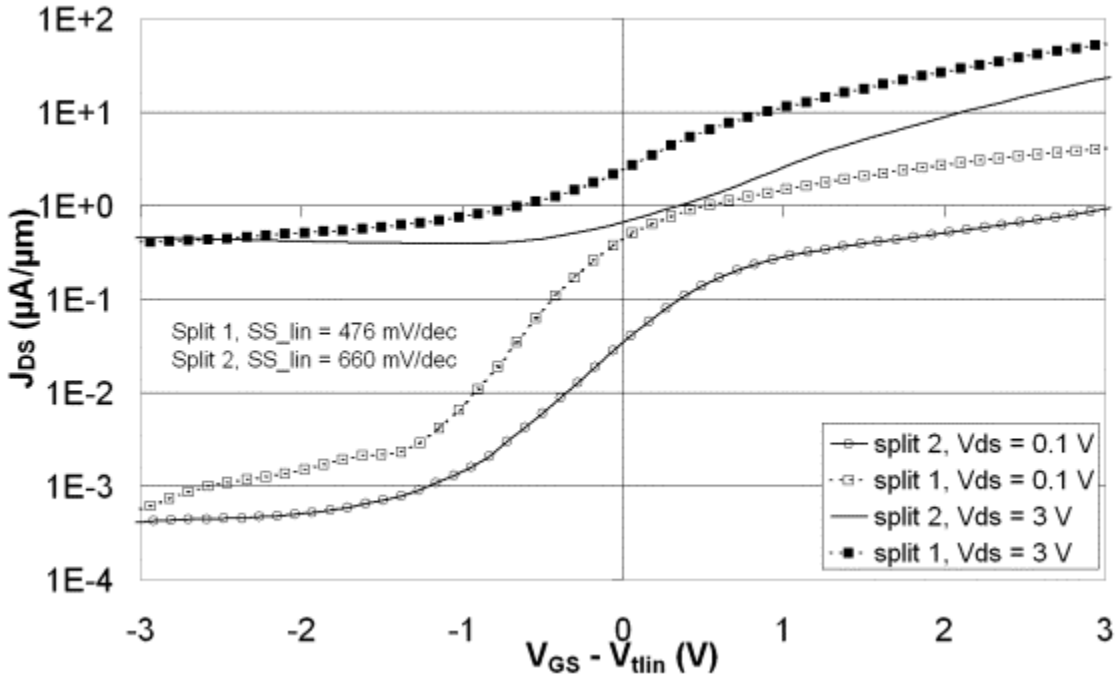

Fig. 7.14. NFET transfer curves for splits 1 and 2. $L_{g,m}$ = 2 µm and W_m = 1 µm.

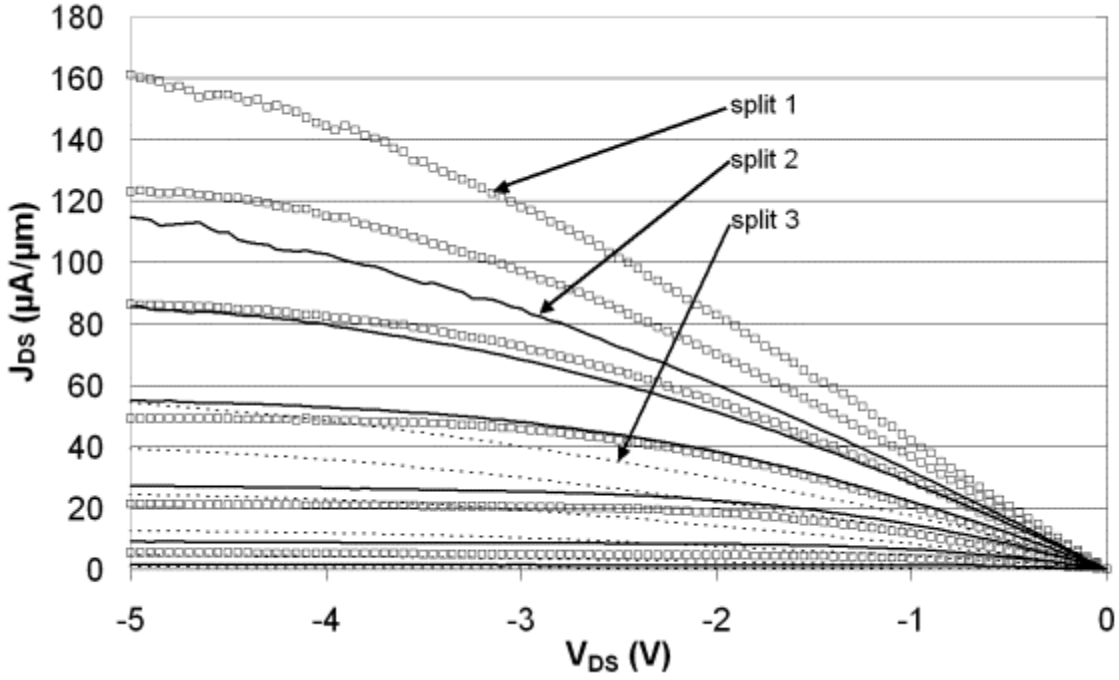

Fig. 7.15. PFET J_{DS} vs. V_{DS} for splits 1-3. $L_{g,m}$ = 2 µm, W_m = 1 µm, and $|V_{GS} - V_{tlin}|$ = 0-5 V in 1 V increments.

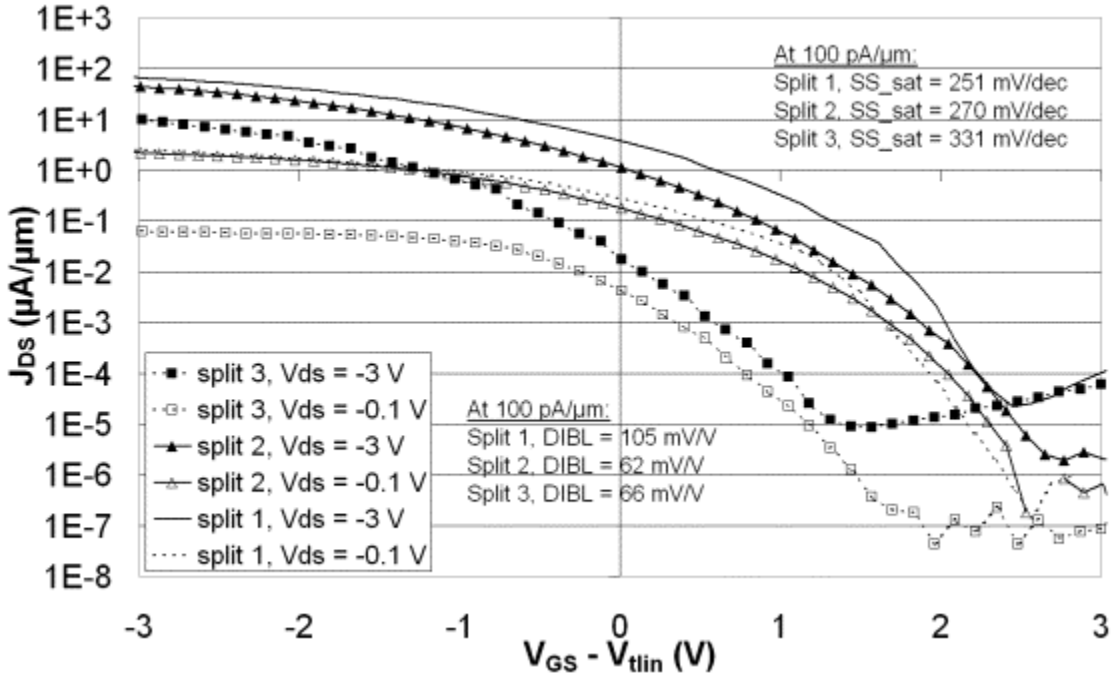

Fig. 7.16. PFET transfer curves for splits 1-3. $L_{g,m}$ = 2 μm and W_m = 1 μm.

Another method of analyzing the effect of the fluorine co-implant is to observe how current scales with $L_{g,m}$. In the ideal case, current scales directly with $1/L_{g,m}$. Thus, for a given $L_{g,m}$, plotting the drive current normalized to the drive current at $L_{g,m}$ = 1 μm should yield a straight line with a slope of 1. Fig. 7.17 shows such a characteristic, where the first and third quadrants are divided into two regions – A and B – and the drive current is taken at $|V_{DS}| = |V_{GS} - V_{tlin}| = 5$ V. If the experimental curve extends into region A, then the slope exceeds 1 and the current scales as $1/(L_{g,m} - \Delta L)$, where ΔL is the effective change in the channel length. This indicates some sort of channel length modulation, and therefore poor SCE immunity. If the experimental curve extends into region B, then the slope is less than 1 and the drive current is limited either by the source/drain series resistance (R_{SD}, the sum of both the source and drain resistances), velocity saturation, and/or the source/drain Schottky barrier. *This is not to state that, in*

region B, the devices exhibit long channel behavior, but rather that other effects are more dominant than a potential emergence of SCE. In the case of the PFET in split 1, current scaling is fairly ideal until $L_{g,m}$ drops below 1 μm, where R_{SD} and/or velocity saturation and/or the Schottky barrier limit current scaling. R_{SD} is removed as a possible factor by Terada-Muta (T-M) measurements, where the PFET R_{SD} was extrapolated to ~ 6.5 kΩ and the NFET R_{SD} is ~ 1.55 kΩ (both from split 1). That the NFET from split 1 exhibits a lower slope in Fig. 7.17, even with a smaller R_{SD}, suggests that R_{SD} is not dominant in this case and that instead the effect of the Schottky barrier and the halo concentration at the M-S interface are the primary components. This notion is further supported by the effect of fluorine on PFET performance being consistent between Fig. 7.17 and Figs. 7.15 and 7.16. It is noted, though, that velocity saturation may also be a significant component in the PFET, due to the linear dependence of J_{DS} on $V_{GS} - V_{tlin}$ in Fig. 7.3.

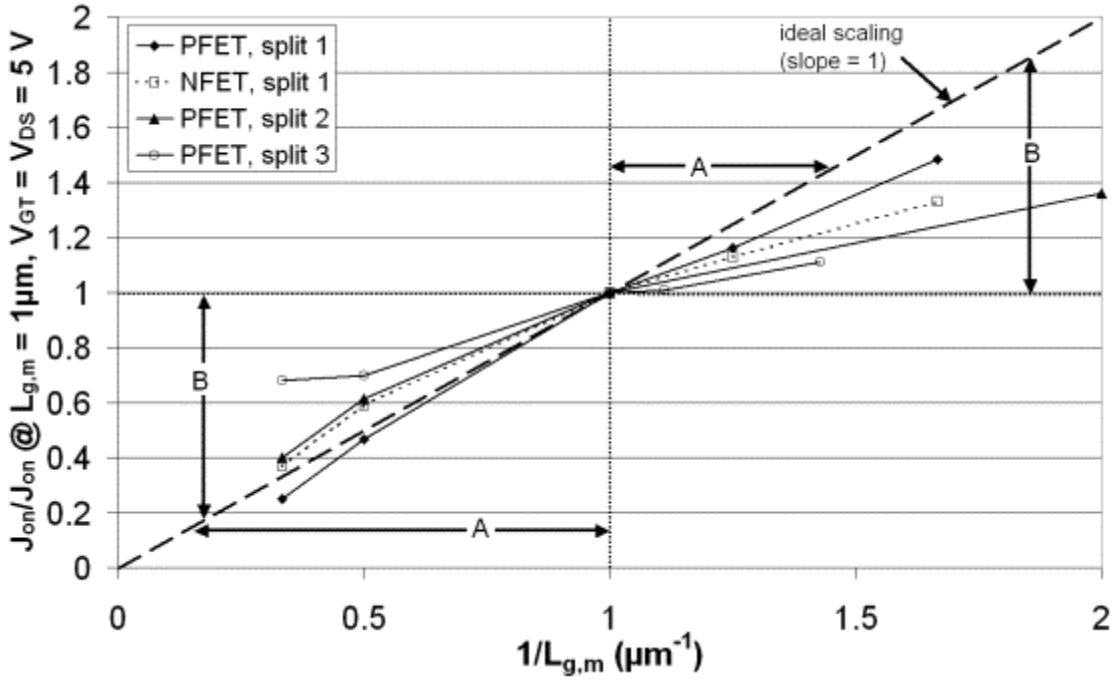

Fig. 7.17. Normalized drive current vs. $1/L_{g,m}$. Region A represents poor SCE immunity and region B represents current scaling limitations due to some effective resistance.

7.6 Band-to-Band Tunneling (BBT) and CMOS Implications

As discussed earlier, and as shown in Fig. 7.8, BBT may limit inverter performance if the body potential sufficiently differs from the source and drain potentials. This can be induced by a gate bias or by an appropriate gate workfunction. Fig. 7.18 shows the inverter VTC for the devices in Fig. 7.8 (anomalous devices from split 2) when the input terminal is driven to voltages sufficient to induce a significant amount of BBT. In such a case, the pull-down and pull-up operations start to reverse. Although, in the case of Fig. 7.18, the V_{in} - V_t swing necessary to show this effect extends beyond the swing expected for V_{DD} = 3 V (±1.5 V), smaller devices, smaller *EOT*, and a sharper halo profile may cause this effect to show up over a smaller range of $V_{in.}$

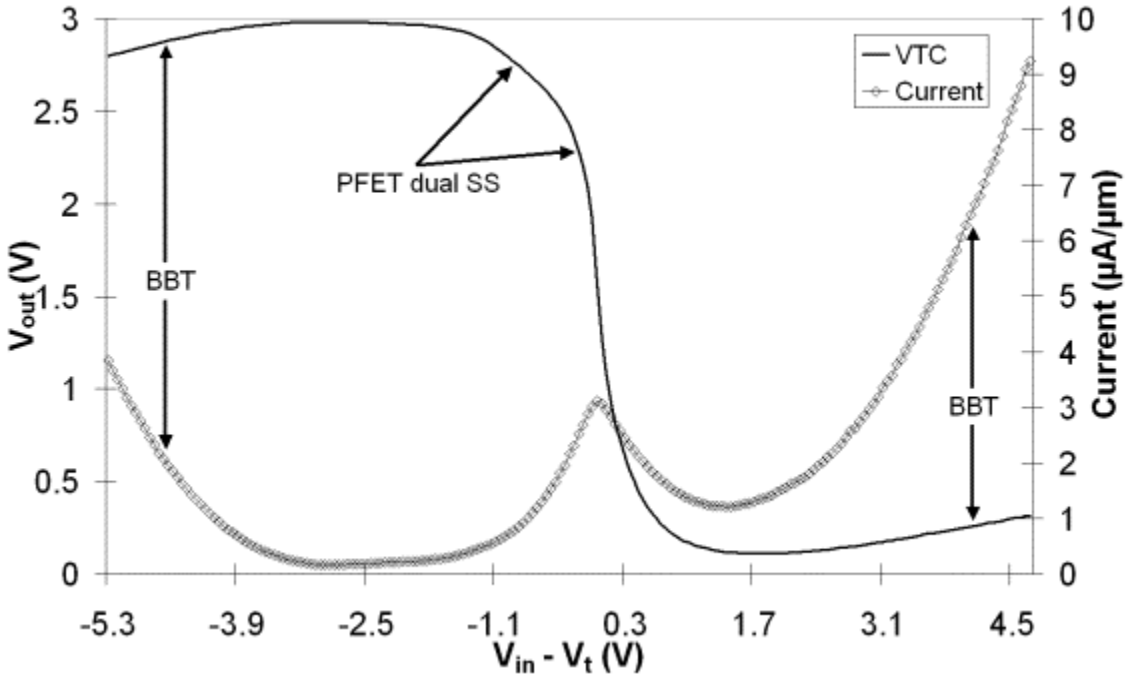

Fig. 7.18. Inverter VTC for devices in Fig. 7.8. V_{DD} = 3 V.

Fig. 7.19 shows the J_{DS} vs. V_{DS} curves in BBT mode for the NFET and PFET from Fig. 7.8, and a number of effects are noticed. First is the diode-like behavior at low

$|V_{DS}|$, which indicates that the drain plays a role in inducing BBT at the source-body junction and/or that tunneling through the halo region itself is a significant current component in this regime of operation. Second is the slope of the "linear" regions for the NFET and PFET. For the PFET, this region is steeper, which suggests that the BBT barrier at the PFET source-body junction is narrower than it is for the NFET for the same $|V_{GS} - V_{tlin}|$ and/or that the tunnel barrier width presented by the halo region is smaller for the PFET. This is supported by Table 7.1 for the W_{halo} results, and can be attributed to the fluorine increasing the halo abruptness for the PFET by counterdoping the tail of the halo profile, while at the same time spreading out the NFET halo region, thus increasing the BBT barrier width for the NFET. This also explains the higher peak current for the PFET. Third is the steepness of the negative differential resistance (NDR) region, which is higher for the NFET than for the PFET. Again, this is attributable to the fluorine counterdoping the p-type halo, but also to the higher solid solubility limit of phosphorus in silicon. In other words, conduction at the *drain* before the NDR region shows up has a higher BBT component (as opposed to tunneling through the halo region) for the NFET than the PFET. This also explains the higher V_{DS} required in the NFET to induce NDR (~ 4 V in the NFET vs. ~ 2.5-3 V in the PFET), as a higher valence-to-conduction band offset requires a higher V_{DS} to reduce V_{GD} such that BBT at the drain starts to cut off.

It is noted that BBT is not observed in all of the NFETs tested; however, this is because most of the NFETs tested exhibited very high subsurface leakage. Looking back at Fig. 7.8, the NFET BBT approaches ~ 0.1 μA/μm at high $V_{GS} - V_{tlin}$, but the subsurface leakage in most NFETs tested (e.g., Fig. 7.14) is approximately 5x larger. It is therefore very reasonable to suggest that BBT does indeed occur in most if not all of the NFET

structures tested, but the higher subsurface leakage prevents one from actually observing said phenomenon in a test environment.

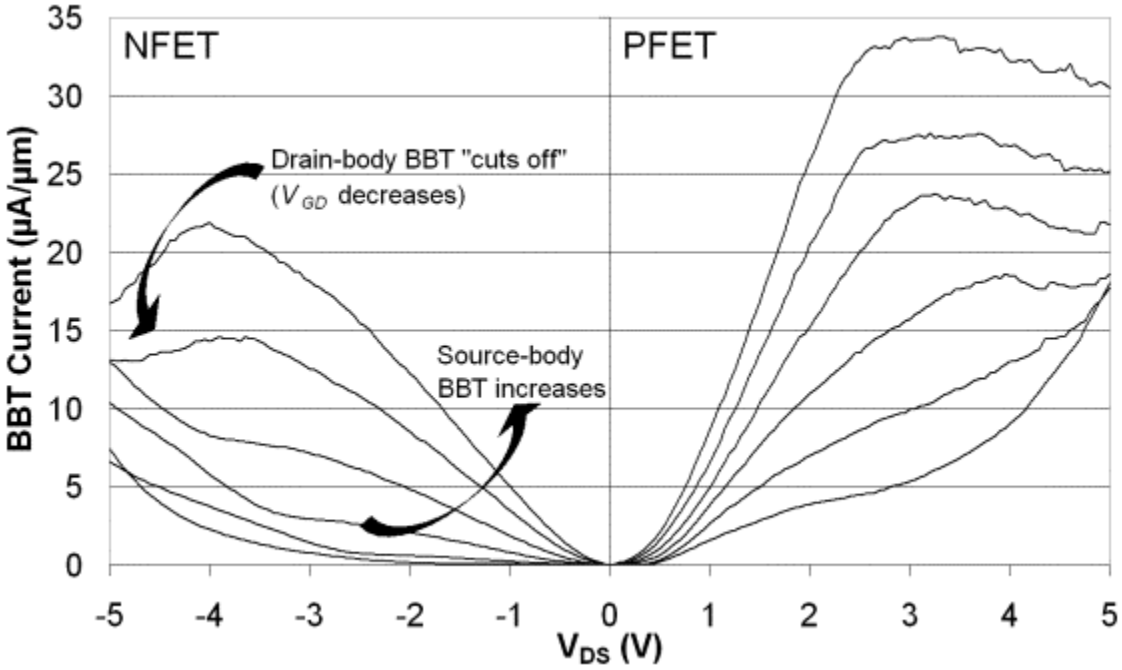

Fig. 7.19. BBT curves for the NFET and PFET from Fig. 7.8. $|V_{GS} - V_{tlin}|$ = 6-11 V in 1 V increments.

7.7 Diode Structures

In Chapter 4, some modeling effort was used to quantify the extent of Schottky barrier lowering (SBL) as a function of N_{halo}, and it was stated that the modeling results were likely an underestimate of the actual extent of SBL for a given N_{halo}. At least in part, this was due to a somewhat arbitrary quantification of the fitting parameter α in the SBL equation, as well as a lack of knowledge on how α may or may not change with N_{halo}. With that model, if the effective SBH is known, one cannot provide an estimate of α without knowing N_{halo}, which cannot be achieved without SIMS analysis. In any case, the test chip for the presented work includes a structure for measuring the effective SBH

for the ITS process utilized in this study. On one end of the structure is a conventional M-S junction (Diode 1 in Fig. 7.20). On the other end is the M-S junction with the ITS-formed halo region, which forms two diodes – one at the M-S junction (Diode 3 in Fig. 7.20) and one between the halo region and the body region (Diode 2 in Fig. 7.20). Since the halo implant only takes place for Diode 3, and therefore Diode 1 is left "bare" during the post-ITS anneal, it is reasonable to assume that Diode 1 will act more as a resistor due to leakage induced by silicide agglomeration during the post-ITS anneal. As such, Diodes 2 and 3 are assumed to dominate the I-V characteristics over a given voltage sweep.

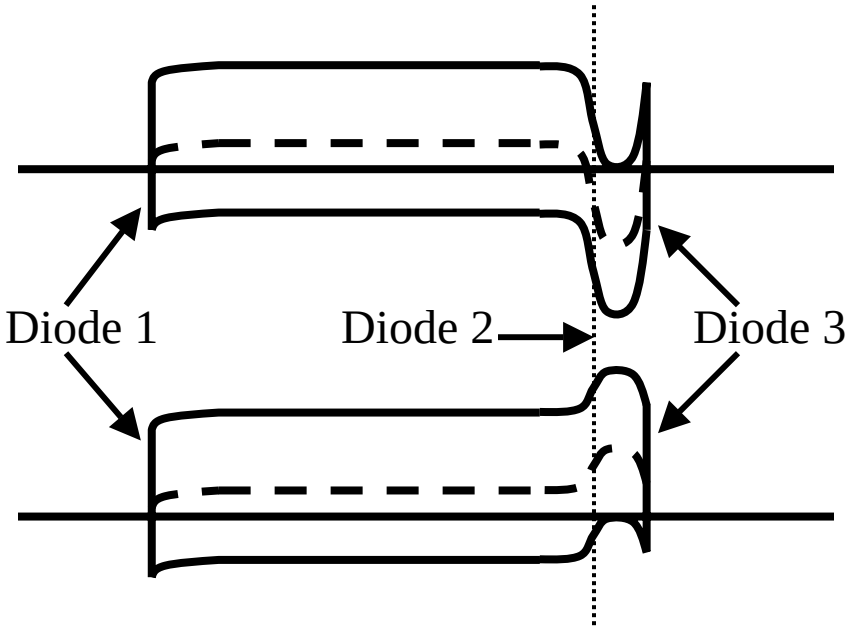

Fig. 7.20. Energy band illustrations for the diode structures used to extract the effective SBH. The top diagram is for the n-type halo region, while the bottom is for the p-type halo region. For both structures, the body region is p-type (4-8x10^{14} cm^{-3}).

For both the n-type and p-type halo structures, the contact to Diode 1 is grounded, while the contact to Diode 3 is swept between ±2 V. Fig. 7.21 shows I-V sweeps for the three process splits performed. For each split, only one diode I-V is attainable (i.e., for

the p-type halo or the n-type halo), but the information in Fig. 7.21 is useful nonetheless. For the n-type halo structures, the halo-body diode turns on with a negative applied bias, while the opposite occurs for the p-type halo structure. Fig. 7.21 shows that the n-type halo from split 1 exhibits the highest threshold voltage (-0.7 V) and the highest subthreshold leakage. This is attributable to the defective M-S junction at Diode 3 (positive V_a), as well as a defective subsurface region in Diode 1, much like how the NFET leakage mentioned previously is higher than the PFET leakage. This is also reflected in the p-type halo diode, which does not exhibit such subthreshold leakage in forward bias (i.e., the p-type halo traverses the entire body depth).

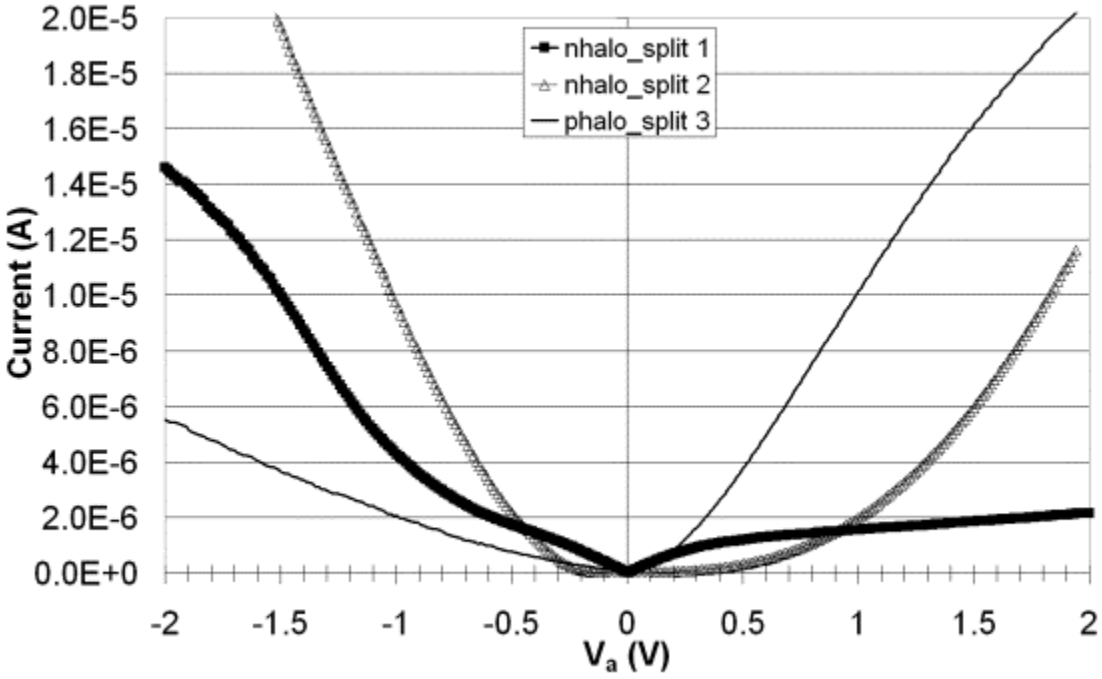

Fig. 7.21. Diode I-V curves for the available n-type and p-type halo structures. For all structures, the body region is p-type (4-8x10^{14} cm^{-3}).

Adding fluorine (split 2) helps the diode leakage considerably by passivating the subsurface region in Diode 3 and by passivating all of Diode 1 (recall, the fluorine

implant was a blanket implant). This also reduces the threshold voltage for the n-type halo-body diode to -0.5 V, which lends further weight to the idea that fluorine reduces phosphorus diffusion within the silicide for a given post-ITS anneal, thus reducing N_{halo}. Also, the reverse leakage characteristics for the n-type halo structure are more diode-like. This suggests that, again, the fluorine is acting as an n-type dopant, resulting in a p-n diode in series with a reverse-biased Schottky diode at Diode 1 as opposed to a leaky M-S junction. This may also explain the higher forward current (negative V_a) for the split 2 vs. split 1 n-type halo regions, where the fluorine creates an n-type region in the subsurface region of Diode 3. Although the dopant concentration at the M-S junction of Diode 3 is presumably lower, the *depth* traverses the entire body thickness, suggesting that the halo depth in the NFETs from split 1 (e.g., Fig. 7.3) is what limits the NFET drive current. If the n-type halo depth were the same as the p-type halo depth for the devices in Fig. 7.3 (i.e., it traverses the entire body thickness), the NFET leakage would not only be lowered considerably, as mentioned previously, but also the NFET drive current would be improved, perhaps on the order of ~ 200 % (assuming the phosphorus diffusion in the source/drain silicide is largely confined to where implant damage exists, per Fig. 7.10). More to this end, if the gate were fully-silicided in the presented structure (i.e., EOT = 9 nm), the PFET drive current at $|V_{DS}| = |V_{GS} - V_{tlin}|$ = 5 V should be on the order of **650 μA/μm**, while the NFET drive current under the same conditions may well exceed **1 mA/μm** for $L_{g,m}$ = 0.6 μm.

Using (2) in Chapter 2, one can use the diode I-V characteristics to estimate the effective SBH. That equation, however, assumes zero voltage dependence on the reverse bias current, which is why the results in Fig. 7.22 are from a relatively low reverse bias of

6.5 mV (other work [12] used higher reverse biases, but with a more complicated model). Even with this in mind, it is doubtful that Fig. 7.22 represents an accurate measurement of the effective SBH at the halo-silicide junction. In fact, Arrhenius measurements (J_R/T^2 vs. $1000/T$) revealed, in several cases, a *positive* slope as opposed to a negative slope, and in many other cases a very jagged plot that is more or less devoid of useful information for extracting the SBH. This indicates that the effective SBH is so small that one would need a temperature range well below 30 °C (on the order of –100 °C [12]) for a reasonably accurate measurement. Even without an "accurate" SBH measurement, what this suggests is that barrier lowering from ITS is substantial. ITS therefore holds considerable promise as a competitor to SIIS [10], [13] for reducing the effective SBH at the M-S junction, even with relatively low temperature post-ITS anneals.

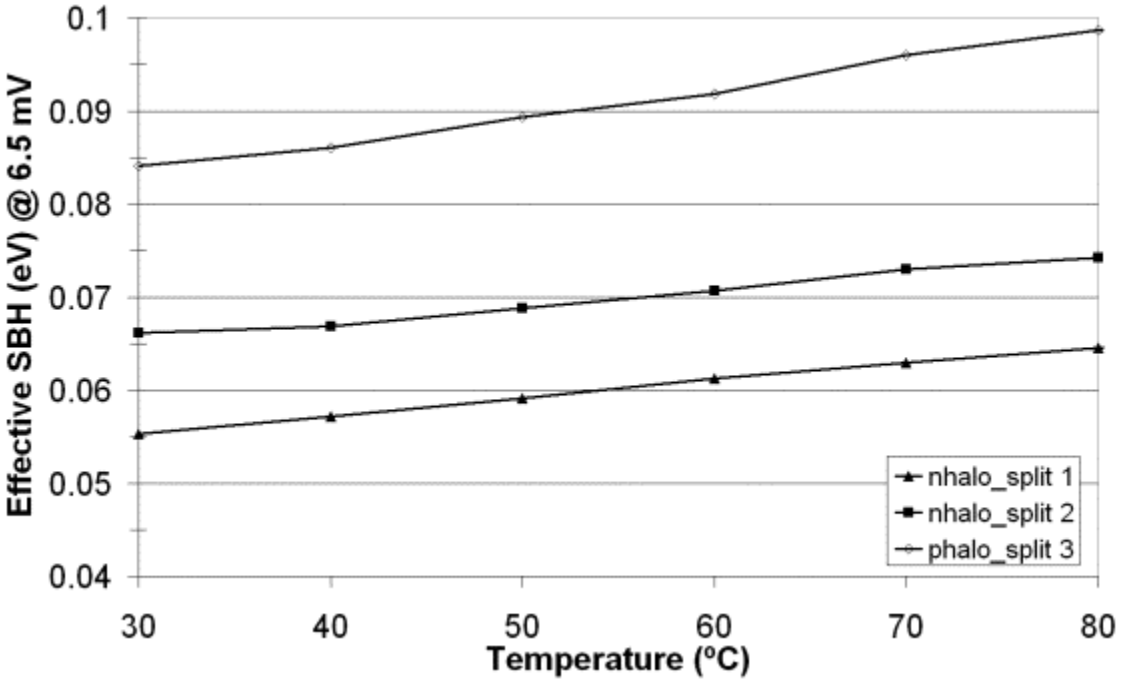

Fig. 7.22. Effective SBH vs. temperature for the available n-type and p-type halo structures. Extracted values are below 0.1 eV at 30 °C, although very low temperature testing would be required to support such a claim.

7.8 Analysis of Potential Counterdoping Effect at the Output Terminal

It was mentioned in Chapter 6 that, since the V_{out} terminal of an inverter is shared by the NFET and PFET in the presented structure, the halo regions at each end of the terminal may be counterdoped by the opposite dopant type diffusing from the other end of the terminal, thus increasing series resistance. For the inverter structures tested, the mask-defined gate spacing (i.e., the width of the V_{out} terminal) is 5 μm. As the n-type and p-type halo implant windows were designed to consume half of that spacing each, the implanted dopants would have to diffuse 2.5 μm to begin counterdoping the opposing halo region. For a 600 °C, 30 min. post-ITS anneal, this is highly unlikely, and is reflected in the fact that performance asymmetry was not observed for most of the samples tested. As fluorine seems to limit boron diffusion within NiSi due to interstitial blocking/competition, it is reasonable to suggest that phosphorus blocks boron diffusion within NiSi in a similar fashion. However, this claim can only be substantiated or refuted at very aggressive scales, where the gate spacing is on the order of the sidewall spacer width used in this study (10's of nm).

Fig. 7.23 shows one of the few cases where performance asymmetry was measured. In the "forward" convention, the V_{out} terminal was treated as the drain for both transistors, while the "reverse" convention used the ground and V_{DD} terminals as the drain for the NFET and PFET, respectively. The observed asymmetry is more pronounced for the NFET, where there exists a sub-linear region for the reverse convention and much lower drive current. This may suggest that boron diffuses much faster in NiSi than phosphorus, and in this case to the point where the n-type halo is counterdoped enough to make the role of the Schottky barrier on current injection

significant again. It is quite possible that the asymmetry observed for the PFET (which only shows up at high V_{GS}) is merely a testing artifact, as it was found that running the same I-V sweep multiple times results in some slight variation between iterations. This may be attributable to non-ideal contact between the aluminum and the NiSi (i.e., the aluminum liftoff process was not very clean, leaving residue on the wafer, and so did not permit a sinter to be performed afterwards). The boron counterdoping theory, again, though, is questionable, due to the size of the gate spacing and the limited number of samples that exhibited this behavior. It is perhaps more likely that the n-type halo implant window was not wide enough (i.e., underexposed during lithography) to open up enough of the V_{out} terminal, thus limiting how much phosphorus was actually implanted into this region.

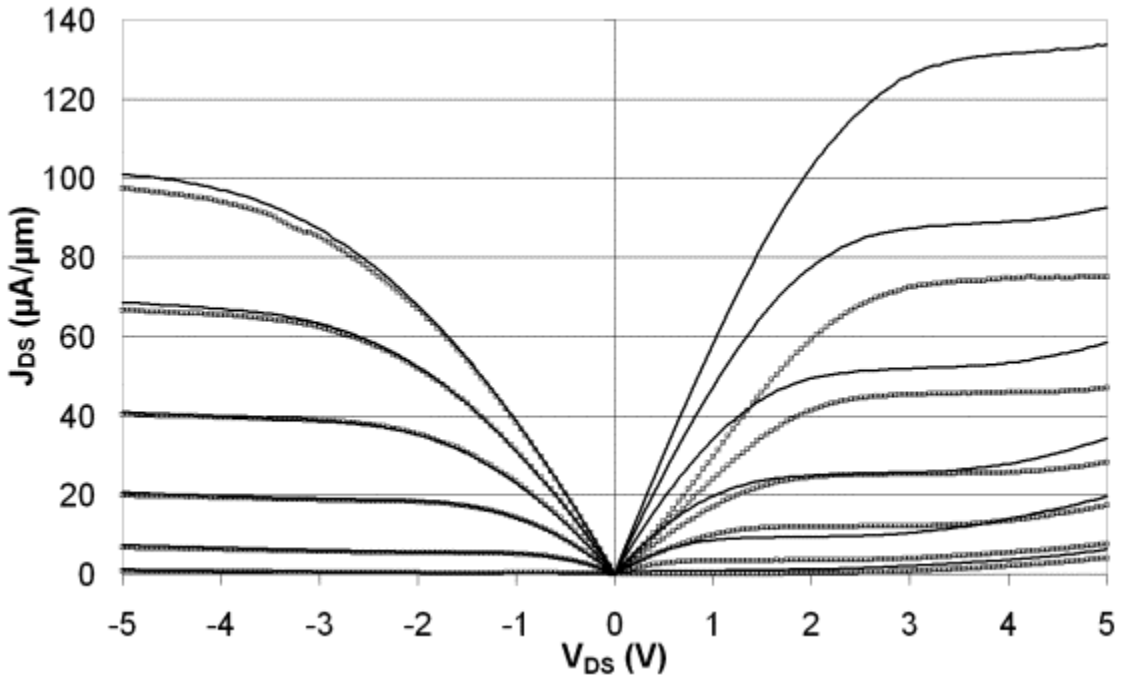

Fig. 7.23. Demonstration of performance asymmetry from one sample in split 1. $L_{g,m}$ = 2 μm, W_m = 1 μm, and $|V_{GS} - V_{tlin}|$ = 0-5 V in 1 V increments. Solid lines represent the forward convention, while the open squares represent the reverse convention.

Chapter 7 References

[1] B.-Y. Tsui, C.-.P. Lin, "A Novel 25-nm Modified Schottky Barrier FinFET With High Performance," *IEEE Elec. Dev. Lett.*, Vol. 25, no. 6, 2004, pp. 430-432.

[2] B.-Y. Tsui, C.-P. Lin, "Process and Characteristics of Modified Schottky Barrier (MSB) p-channel FinFETs," *IEEE Trans. Elec. Dev.*, Vol. 52, no. 11, 2005, pp. 2455-2462.

[3] C.-P. Lin, B.-Y. Tsui, "Hot-Carrier Effects in P-Channel Modified Schottky-Barrier FinFETs," *IEEE Elec. Dev. Lett.*, Vol. 26, no. 6, 2005, pp. 394-396.

[4] C.-P. Lin, B.-Y. Tsui, "Characteristics of Modified-Schottky-Barrier (MSB) FinFETs," *VLSI Tech.*, 2005, pp. 118-119.

[5] C.-F. Huang, B.-Y. Tsui, "Short-Channel Metal-Gate TFTs with Modified Schottky-Barrier Source/Drain," *IEEE Elec. Dev. Lett.*, Vol. 27, no. 1, 2006, pp. 43-45.

[6] C.-C. Wang, C.-J. Lin, M.-C. Chen, "Formation of NiSi-Silicided p^+n Shallow Junctions Using Implant-Through-Silicide and Low-Temperature Furnace Annealing," *J. Electrochem. Soc.*, Vol. 150, no. 9, 2003, pp. 557-562.

[7] F. Deng, K. Ring, Z.F. Guan, S.S. Lau, W.B. Dubbelday, N. Wang, K.K. Fung, "Structural investigation of self-aligned silicidation on separation by implantation oxygen," *J. App. Phys.*, Vol. 81, no. 12, 1997, pp. 8040-8046.

[8] Q.T. Zhao, E. Rije, U. Bruer, St. Lenk, S. Mantl, "Tuning of Silicide SBHs by Segregation of Sulfur Atoms," *Proc. IEEE*, 2004, pp. 456-459.

[9] R.F. Pierret, "Semiconductor Device Fundamentals," *Addison-Wesley Publishing Company, Inc.*, 1996, p. 493.

[10] F. Deng, R.A. Johnson, P.M. Asbeck, S.S. Lau, W.B. Dubbelday, T. Hsiao, J. Woo, "Salicidation process using NiSi and its device application," *J. App. Phys.*, Vol. 81, no. 12, 1997, pp. 8047-8051.

[11] J. Kedzierski, D. Boyd, C. Cabral, Jr., P. Ronsheim, S. Zafar, P. M. Kozlowski, J. A. Ott, M. Ieong, "Threshold Voltage Control in NiSi-Gated MOSFETs Through SIIS," *IEEE Trans. Elec. Dev.*, Vol. 52, no. 1, 2005, pp. 39-46.

[12] E. Dubois, G. Larrieu, "Measurement of low Schottky barrier heights applied to metallic source/drain metal-oxide-semiconductor field effect transistors," *J. Appl. Phys.*, Vol. 96, no. 1, 2004, pp. 729-737.

[13] A. Kinoshita, Y. Tsuchiya, A. Yagashita, K. Uchida, J. Koga, “Solution for high-performance Schottky-source/drain MOSFETs: Schottky barrier height engineering with dopant segregation technique,” *VLSI Symp. Tech. Dig.*, 2004, pp. 168-169.

Chapter 8

Negative Differential Resistance (NDR) in Conventional SFETs

8.1 Observance of NDR in This and Other Work

The author's first observance of the NDR characteristic in conventional SFETs was in the prototype devices fabricated for the undergraduate senior design project during Winter/Spring 2004. For said devices, the source/drain regions (and the gate region) were pure aluminum, the gate dielectric was 500 Å of Tetraethyl Orthosilicate (TEOS), and the body region was lightly doped n-type (bulk substrate, 5-15 Ω-cm). Although far from ideal from a performance standpoint, it was sufficient to demonstrate the concept of Schottky source/drain regions. However, the n-channel J_{DS} vs. V_{DS} curves consistently exhibited NDR characteristics, some examples of which are shown in Fig. 8.1.

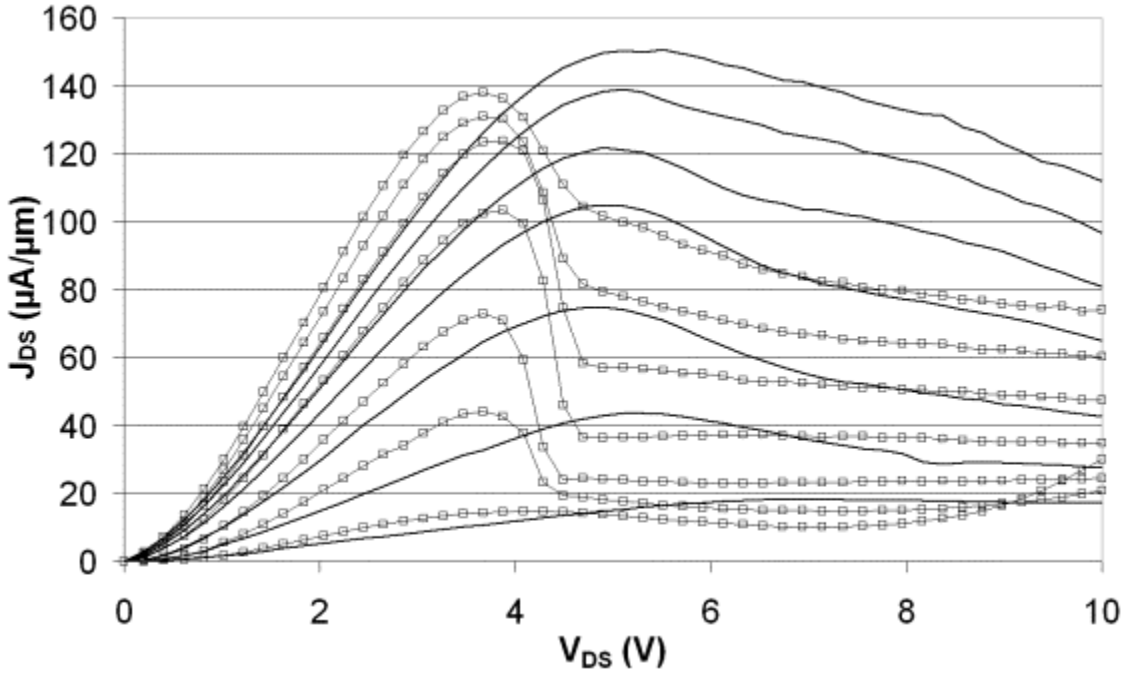

Fig. 8.1. NDR characteristic observed in prototype SFETs from the senior design project in 2004. V_{GS} = 4-10 V in 1 V increments and the gate length L_G ~ 4.1 μm after etching. The solid lines and the open box lines represent two different transistors.

As the results in Fig. 8.1 are from devices with unmodified Schottky barriers (i.e., no halo regions), it is interesting that NDR is observed if one assumes an ideal M-S junction and ideal SFET operation (Chapter 3). Clearly, however, this is not the case. Such a characteristic has also been observed in at least one other body of work [1], shown in Fig. 8.2 for $V_g - V_{th}$ = 5 V and V_{DS} ~ 2.8 V. In this device, the source/drain regions were $CoSi_2$ with a p-type body region (bulk silicon) at $1x10^{17}$ cm^{-3}, and L_G = 0.4 μm.

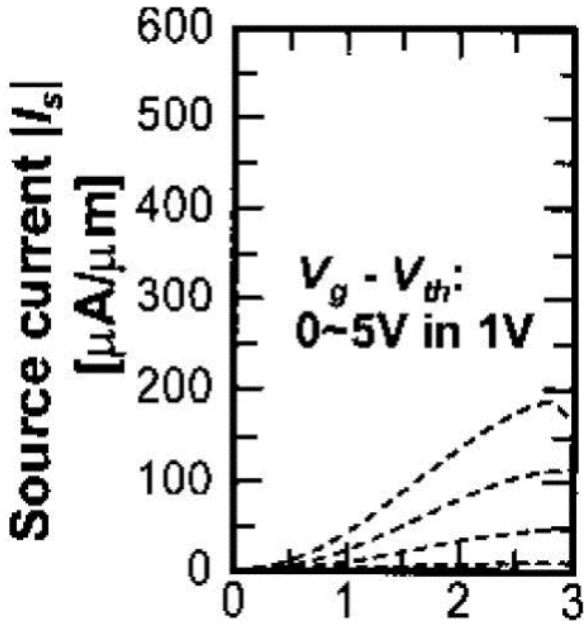

Fig. 8.2. NDR characteristic observed in other work [1] at high $V_g - V_{th}$ and high V_{DS}.

In the presented work, on one wafer, the yielding NFETs were ambipolar, indicating that the implanted phosphorus did not diffuse to the interface (perhaps due to temperature variation in the RTA system). Interestingly, these NFETs also exhibited NDR characteristics during n-channel operation, an example of which is shown in Fig. 8.3. That NDR has been observed with NiSi, $CoSi_2$, and Al source/drain regions in different device structures suggests that NDR does not have a structural or material dependence so much as it is a function of the nature of the M-S interface.

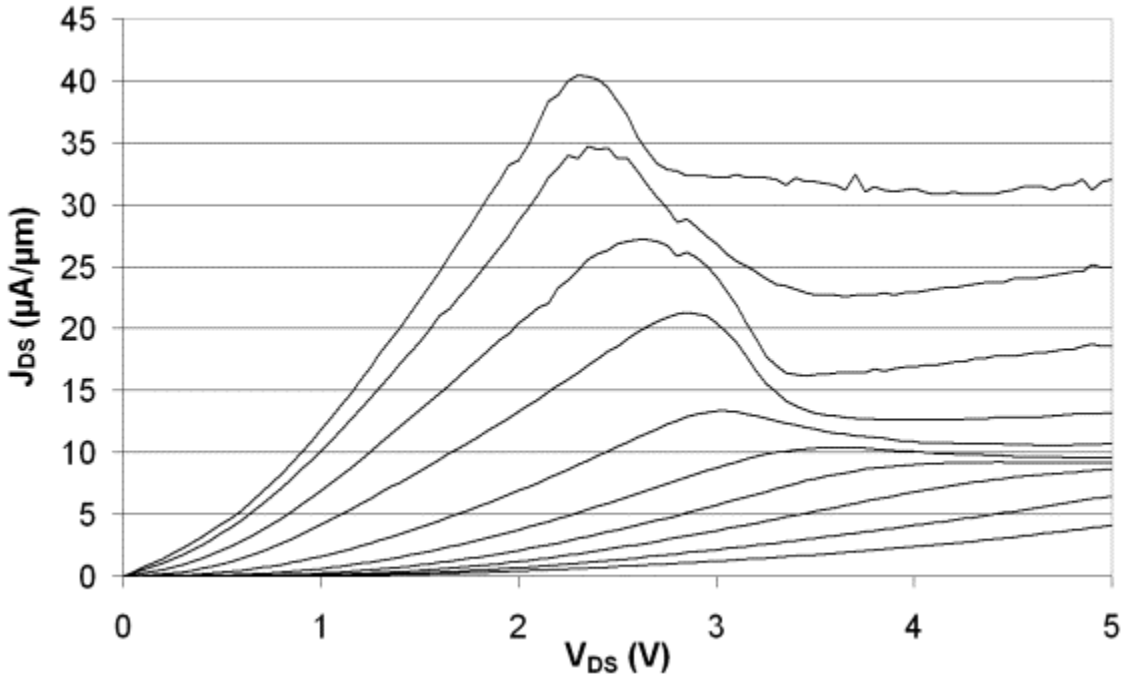

Fig. 8.3. NDR characteristic observed in this work for an NFET yielding ambipolar behavior. $L_{g,m}$ = 3 μm, W_m = 1 μm, and V_{GS} = 0-9 V in 1 V increments.

8.2 Proposed Physical Mechanism

It is proposed that the primary physical mechanism behind the observed NDR characteristic is band-to-band tunneling (BBT) due to a distribution of acceptor-like states as a function of energy, as well as distance from the M-S junction. In some cases, it also seems that BBT is trap-assisted. As the gate is driven to higher positive values, more of these acceptor-like states fill up with electrons, effectively "pinning" the energy bands at and near the M-S junction to some potential profile. This results in the gate losing control of the channel potential in these occupied regions, in some sense forming an artificial halo region similar to what was formed chemically in Chapter 7. As the potential of the remainder of the channel is still under gate control, eventually a valence-to-conduction band overlap between this region and the "halo" region sufficient to induce BBT at the source and drain occurs. As V_{DS} is increased, current increases at first, but

eventually BBT is "cut off" as the gate-to-drain bias, V_{GD}, decreases. At this point, the NDR region is visible, and eventually the current flattens out at high V_{DS} due to the relative independence of BBT at the source on V_{DS}. This is illustrated in Fig. 8.4.

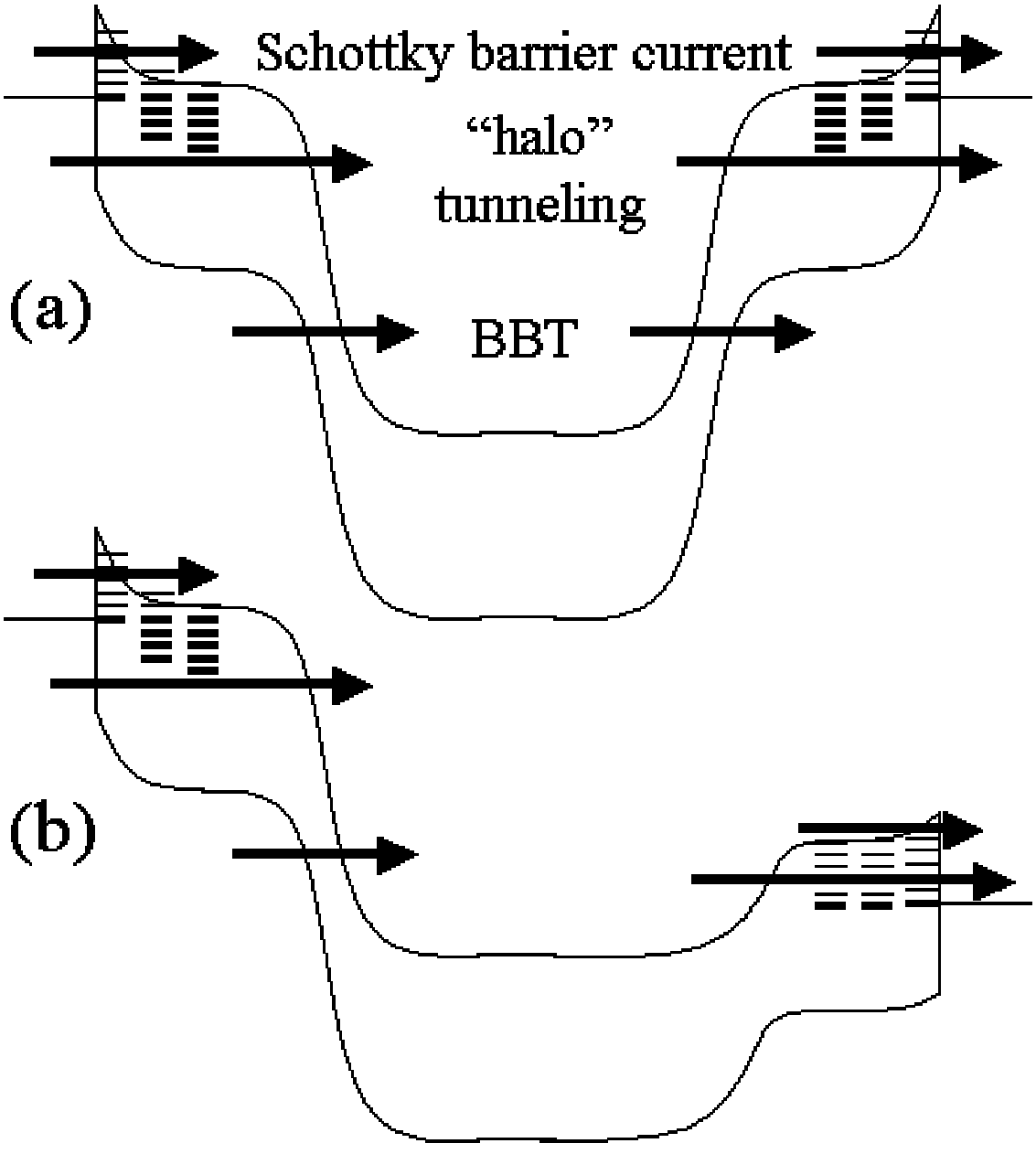

Fig. 8.4. Band diagrams illustrating NDR characteristic in conventional SFETs. In (a), interface states fill up (bold dashes) at progressively higher V_{GS}, facilitating BBT at the source and drain. In (b), as V_{DS} increases, BBT at the drain cuts off and some interface states empty out (narrow dashes); conduction at the drain is reduced to the Schottky barrier transmission and "halo" tunneling components.

Depending on the nature of the interface states, the existence of said states can either increase or decrease current. Fig. 8.5 outlines a "burn-in" process for one of the SFETs from the senior design project. In this device, V_{GS} was held at 10 V while V_{DS} was repeatedly swept from 0-10 V. For each iteration, the I_{DS} vs. V_{DS} characteristics changed. At moderate V_{DS}, where the NDR region is initially very sharp, the peak current drops off with each iteration. This indicates that BBT in this particular device is assisted by traps within the energy band that exhibit a relatively high occupation time, and as these traps fill up, coulombic repulsion reduces the BBT component of the total current. In the high V_{DS} regime, the current increases for the first few iterations and then decreases slightly. It is possible that, as the traps that assist in BBT fill up, secondary traps with a relatively low occupation time are formed which assist in tunneling through the "halo" regions.

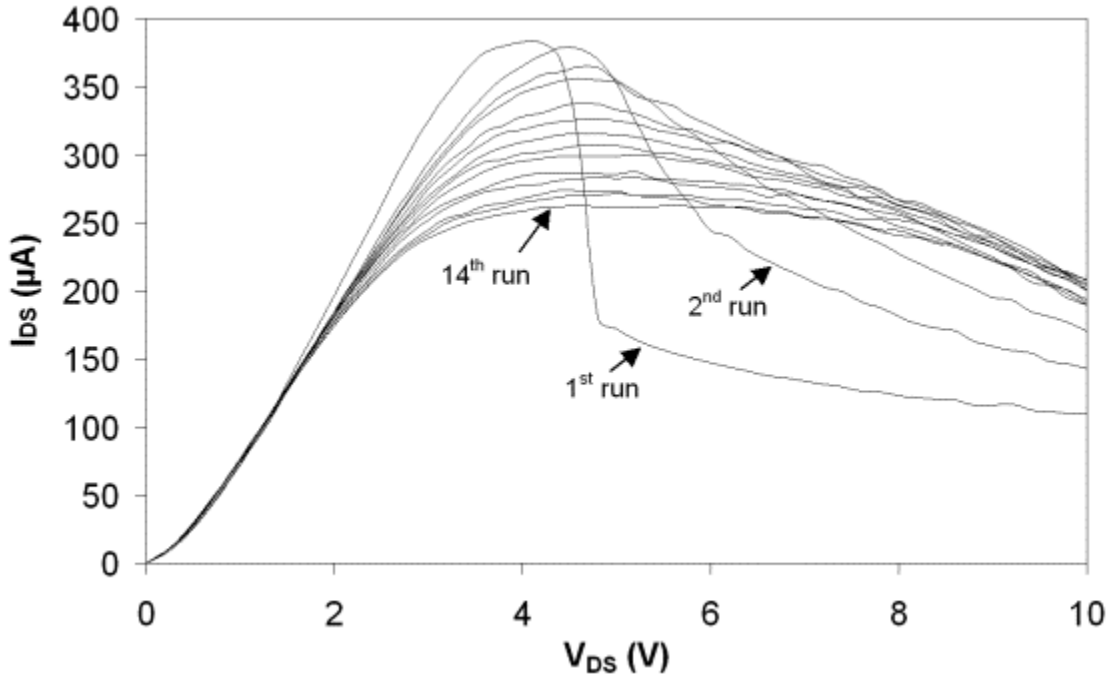

Fig. 8.5. Example burn-in characteristics for an aluminum source/drain SFET from the senior design project. L_G ~ 4.1 μm.

For the NiSi source/drain SFETs fabricated in this study which exhibited ambipolar behavior and NDR characteristics, performing a burn-in like that in Fig. 8.5 yielded very little if any change in the I-V curve. That is, the peak current at moderate V_{DS} did not reduce and the saturation current at high V_{DS} did not change either. That NDR was still observed (Fig. 8.3) suggests that the traps assisting in BBT in these particular devices have an occupation time that is too low to cause sufficient charge buildup to reduce BBT. This may suggest a relationship between trap occupation time and whether the M-S junction was formed by silicidation or simply by metal deposition.

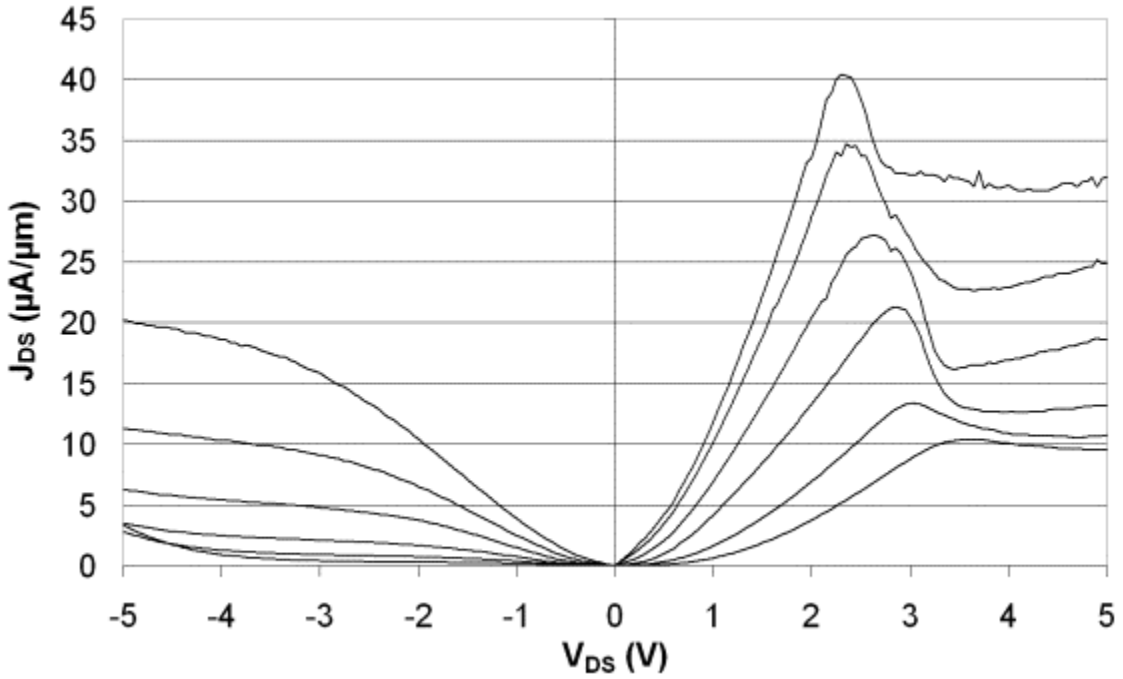

Fig. 8.6. N-channel and p-channel J_{DS} vs. V_{DS} for the device in Fig. 8.3. $|V_{GS}|$ = 4-9 V in 1 V increments.

It is interesting that NDR has not been observed in p-channel operation for both the aluminum and NiSi source/drain SFETs, which would have indicated a high concentration of *donor-like* states at and near the M-S interface. Instead, in the NiSi device, p-channel operation looks like what is expected of a conventional SFET

(Fig. 8.6), whose distinguishing characteristic is a “sublinear” region, otherwise known as rectifying behavior in the linear region. As the Schottky barrier is not reduced by a chemically formed halo region (Chapter 7), but instead must be modulated by the gate, the drive current is considerably lower, as expected. For the aluminum source/drain device, however, p-channel operation looks like a bulk switching device, whereby no sublinear region is observable (Fig. 8.7). As the curve in Fig. 8.7 was obtained after a burn-in process, and as the p-channel curve in Fig. 8.6 exhibits rectifying behavior (for which no burn-in was possible), it is likely that the trap-induced halo region for the aluminum source/drain devices temporarily turned said devices into p-channel bulk switching SFETs due to the higher trap occupation time.

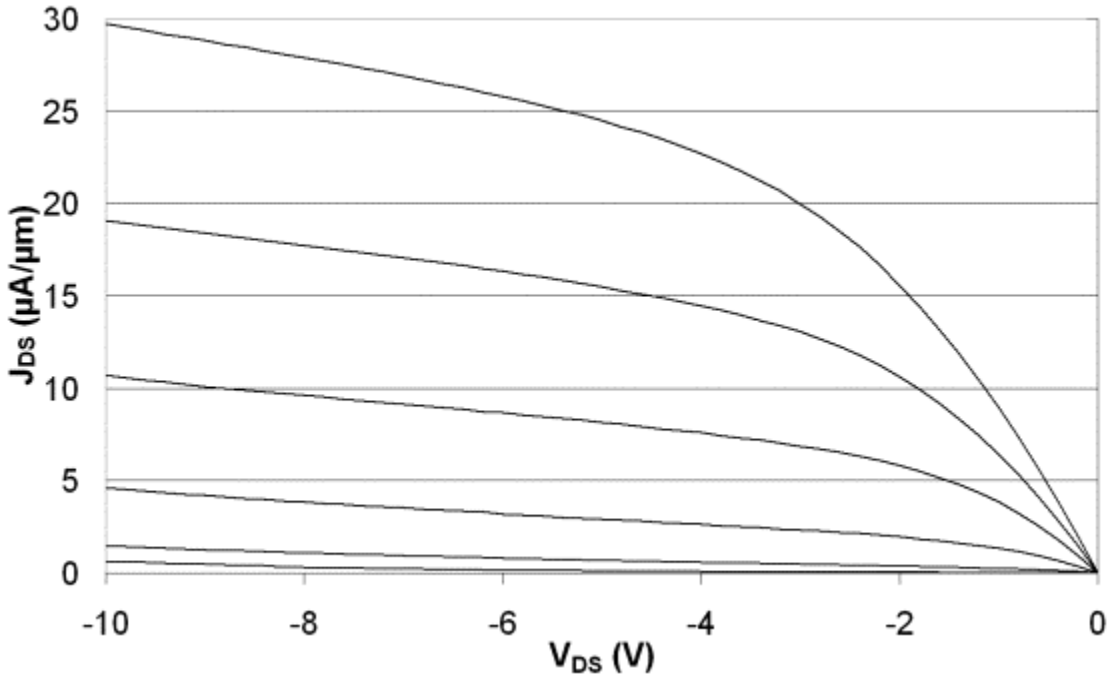

Fig. 8.7. P-channel J_{DS} vs. V_{DS} for an aluminum source/drain SFET from the senior design project. $|V_{GS}|$ = 5-10 V in 1 V increments and L_G ~ 4.1 μm.

To investigate the claim of high lifetime trap states enhancing current injection in p-channel mode for the aluminum source/drain SFETs, more testing was necessary – no

pre-burn p-channel data were collected in 2004 when the devices were initially tested (Figs. 8.1, 8.5, and 8.7 are data from 2004). Going back to test the devices again two years later (for pre- and post-burn data), however, revealed something very interesting – performing a burn-in consistently yielded little if any change in the burn-in I-V curve. That is, instead of something like Fig. 8.5, each successive curve had the same shape as the first iteration in Fig. 8.5, but at much lower current and with little if any increase in current per iteration. This suggests that, over the course of time (be it a few months or the ~2 year gap between the 2004 testing and the 2006 testing), the high occupation time traps were somehow neutralized or reduced in concentration. Also, the measured n-channel current was considerably lower than what had normally been achieved during testing in 2004 (e.g., Fig. 8.1), and so the original ambipolarity of the device was substantially reduced over time. An example of this is shown in Fig. 8.8.

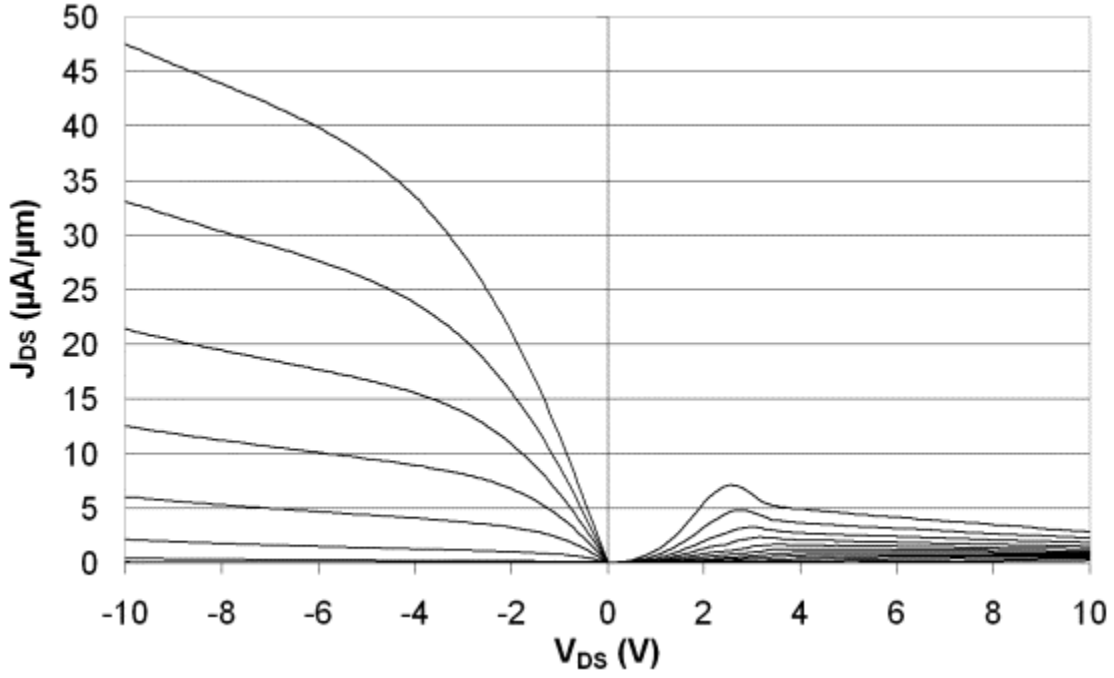

Fig. 8.8. J_{DS} vs. V_{DS} for an aluminum source/drain SFET from the senior design project, tested in March 2006. P-channel $|V_{GS}|$ = 1-10 V in 1 V increments, n-channel V_{GS} = 7-17 V in 1 V increments, and L_G ~ 4.1 μm.

That the NDR characteristics observed for n-channel operation in Fig. 8.8 require considerably higher V_{GS} than what is observed in Fig. 8.1 supports the idea of a reduced acceptor-like state concentration, and suggests that the remaining states are distributed very close to the conduction band. The p-channel current exhibits little if any sublinear response (quite the opposite for n-channel behavior), indicating a very low hole SBH and a very high electron SBH. This is also supported by Fig. 8.9, which shows very little ambipolar leakage, and the observed increase in leakage with $|V_{DS}|$ is attributable to subsurface DIBL. Although the "new" and "old" test results are not from the same exact device, the characteristics of the "new" and "old" results were found to be consistent between multiple samples.

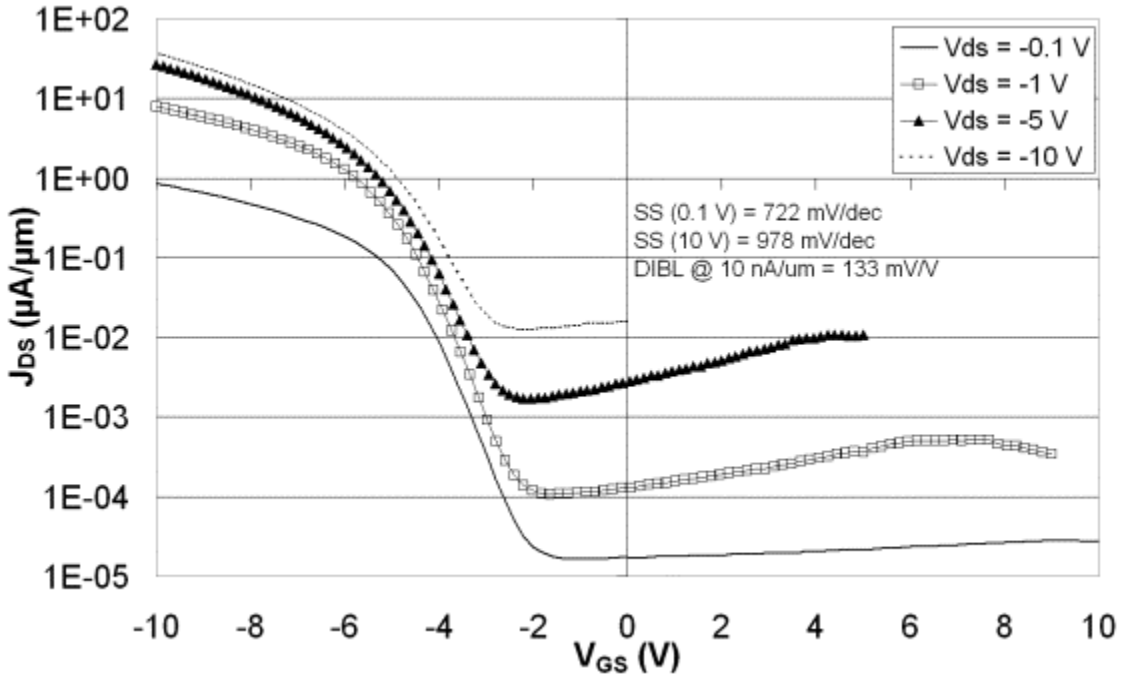

Fig. 8.9. J_{DS} vs. V_{GS} for the aluminum source/drain SFET in Fig. 8.8, tested in March 2006.

Chapter 8 References

[1] A. Kinoshita, Y. Tsuchiya, A. Yagashita, K. Uchida, J. Koga, “Solution for high-performance Schottky-source/drain MOSFETs: Schottky barrier height engineering with dopant segregation technique,” *VLSI Symp. Tech. Dig.*, 2004, pp. 168-169.

Chapter 9

Polysilicon-on-Insulator (POI) SFETs and CMOS Implementation

9.1 Method of Fabrication

The process flow started with p-type (boron-doped, 25-45 Ω-cm, 3-6x10^{14} cm^{-3}) bulk silicon wafers. A 100 nm wet oxide was grown to define the buried oxide (BOX), after which 220 nm of polysilicon was deposited via LPCVD. The wafers were then furnace annealed at 1100 °C for 1 h in N_2, the purpose of which was to reduce the interface charge between the BOX and the body. After defining the active regions, a 35 nm gate oxide was thermally grown, after which 220 nm of polysilicon with a 110 nm nitride cap were deposited via LPCVD. After gate patterning, a 30 nm thick oxide sidewall spacer was grown. The oxide over the source/drain regions was then removed in a dry etch with CHF_3 and O_2 and the nitride cap was stripped in phosphoric acid at 175 °C.

A 30 s, 50:1 HF dip, followed by a 1 min rinse in DI water and then a spin rinse/dry, was performed. The wafers were immediately loaded into a sputter chamber and placed under vacuum. Nickel was then sputter deposited to a target thickness of 120 nm after reaching a base pressure of 1-2 μtorr. The silicidation step was performed at 500 °C for 1 min in N_2 via RTA, and unreacted nickel was removed in a 2:1 H_2O_2:H_2SO_4 mixture at 90 °C.

An ITS process was performed for both the NFETs (phosphorus implant) and the PFETs (BF_2 implant). For both implants, the dose and energy were 4x10^{15} cm^{-2} and

80 keV, respectively. To form the halo regions, a subsequent thermal anneal was performed via RTA at 600 °C for 5 min. After the halo formation, aluminum metallization was performed with an evaporation/liftoff process. A top-down picture of the circuit structure (inverter) before metallization is shown in Fig. 9.1, which shows the NFET and PFET sharing a metallic source/drain (MSD) region at the V_{out} terminal, as discussed in Chapter 6. Like in Chapter 7, the u-shaped structure is the V_{in} terminal, and the terminals to the far right and left are the ground and V_{DD} terminals, respectively. The terminal in the center is the V_{out} terminal.

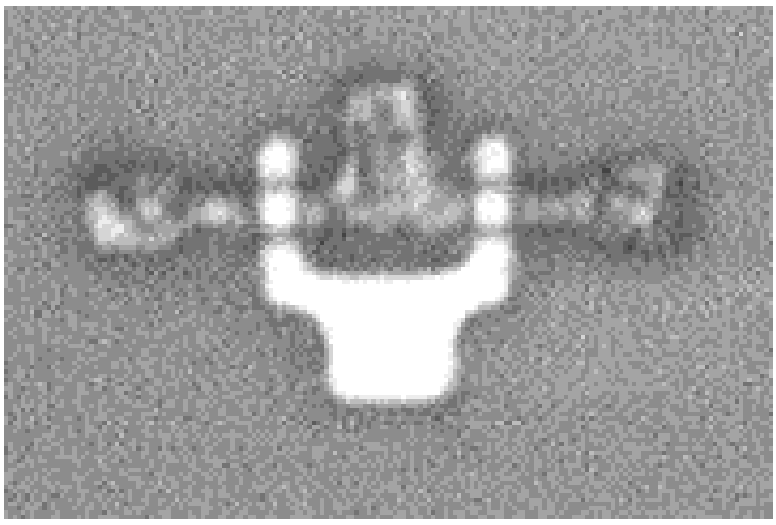

Fig. 9.1. Top-down picture of MSD CMOS inverter before metallization, using a POI substrate. The mask-defined gate length and width are both 1 μm.

Although both the gate and active regions in this structure are polysilicon of the same thickness, and although they should both be fully silicided (according to the as-deposited nickel thickness), it is very interesting to note the different color/tone between the gate and active regions. The gate region looks light, smooth, and uniform (very similar to the silicon active and gate regions shown in Chapter 7), while the active region looks rough and non-uniform. This is also shown in Fig. 9.2, which shows a portion of

the gate-to-active alignment verniers. On the left of Fig. 9.2 is the gate level, while the active level is on the right side. The lighter regions in the active level in Fig. 9.2 are actually green-ish in color, while the darker regions are simply a darker shade of the gate level. As the only difference between the two levels is the furnace anneal performed on the active region but not the gate region, it would seem that this furnace anneal is the cause. What constitutes these green-ish regions is currently unkown (possibly thermal nitride formed during the anneal or an artifact of the oxide etch performed before silicidation), but it seems that the gate and active polysilicon layers have different morphologies.

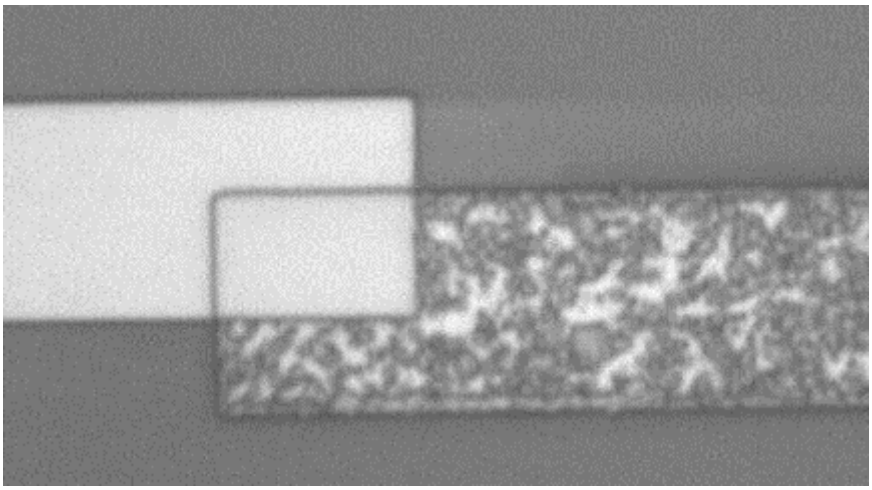

Fig. 9.2. Top-down picture of a portion of the gate-to-active alignment vernier after silicidation. The active level is on the right, while the gate level is on the left.

It is noted that the results that follow by no means represent a diligent study of MSD CMOS on POI substrates. What is presented is purely a proof of concept and a demonstration of the idea that thin film transistors (TFTs) for display applications can be fabricated with the MSD architecture and a low thermal budget (excluding the gate stack in this study, which was formed at relatively high temperatures compared to conventional

TFT fabrication). It is very reasonable to suspect that the electrical results that follow may be considerably improved by, amongst other things, *not* performing said furnace anneal in future studies or by using a different anneal process, such as solid phase crystallization (SPC) at 600 °C for 24 h in N_2 [1], [2].

9.2 Electrical Results for MSD CMOS on POI Substrates

Fig. 9.3 shows the voltage transfer characteristics (VTCs) of a $L_{g,m}$ = 2 μm inverter (NFET:PFET width = 1:1) for V_{DD} = 5, 7.5, and 10 V. As with the SOI results, it is thought that this is the first-ever empirical demonstration of MSD CMOS on POI substrates. Although gain is relatively low compared to the SOI results from Chapter 7 (~ 5.4 V/V at V_{DD} = 10 V for POI vs. ~ 25.7 V/V at V_{DD} = 5 V for SOI), it is to be expected for a thicker (35 nm vs. 18 nm) *EOT* and the lower mobility (which translates to lower transconductance) inherent in polycrystalline body regions.

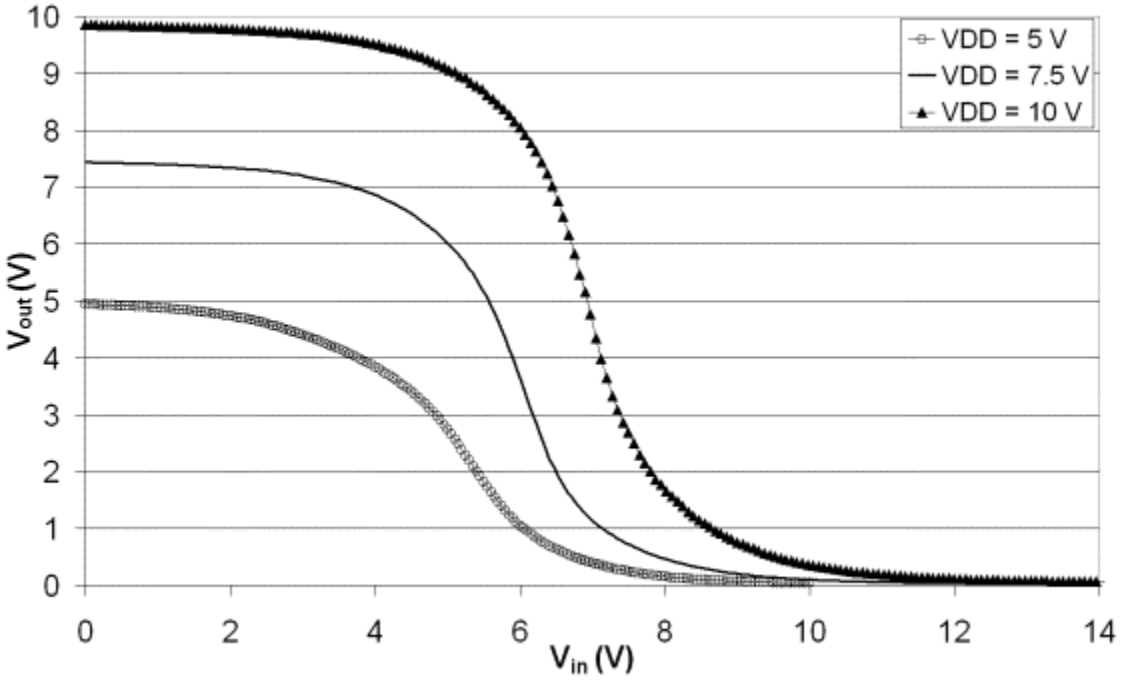

Fig. 9.3. VTCs for MSD CMOS on POI substrates. $L_{g,m}$ = 2 μm, W_m = 1 μm.

It is interesting that the inverter threshold shift is positive, on the order of ~ 2-3 V. Although gate oxide charge is usually a cause of threshold shifts, the occurrence of a positive shift (i.e., negative oxide charge) is very rare. In the case of Fig. 9.3, then, the shift is not likely attributable to oxide charge. This is supported by Fig. 9.4, which shows the J_{DS} vs. V_{DS} characteristics for both the NFET and PFET in Fig. 9.3. The PFET exhibits relatively good behavior, where the trough of the transfer characteristic is placed at V_{GS} ~ 0 V. For the NFET, said trough is at V_{GS} ~ 3.5 V, and the transfer characteristic itself exhibits a dual slope behavior for V_{DS} = 5 V, where the transition between these slopes is at V_{GS} ~ 12.5 V. This indicates that the role of the Schottky barrier in the NFET is more significant than in the PFET, again supporting the notion that phosphorus diffusion in NiSi is slower than boron. It would seem that the inverter threshold shift is therefore attributable to inferior NFET switching, which can be improved with a longer post-ITS anneal and/or a higher phosphorus implant energy.

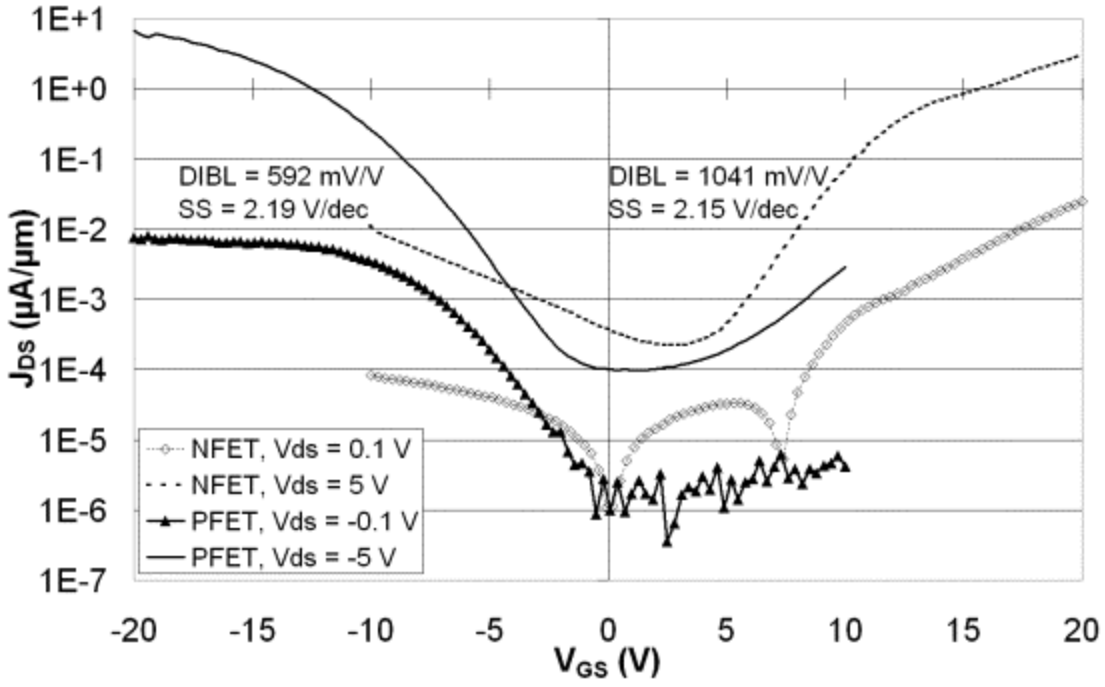

Fig. 9.4. J_{DS} vs. V_{GS} for the NFET and PFET from Fig. 9.3.

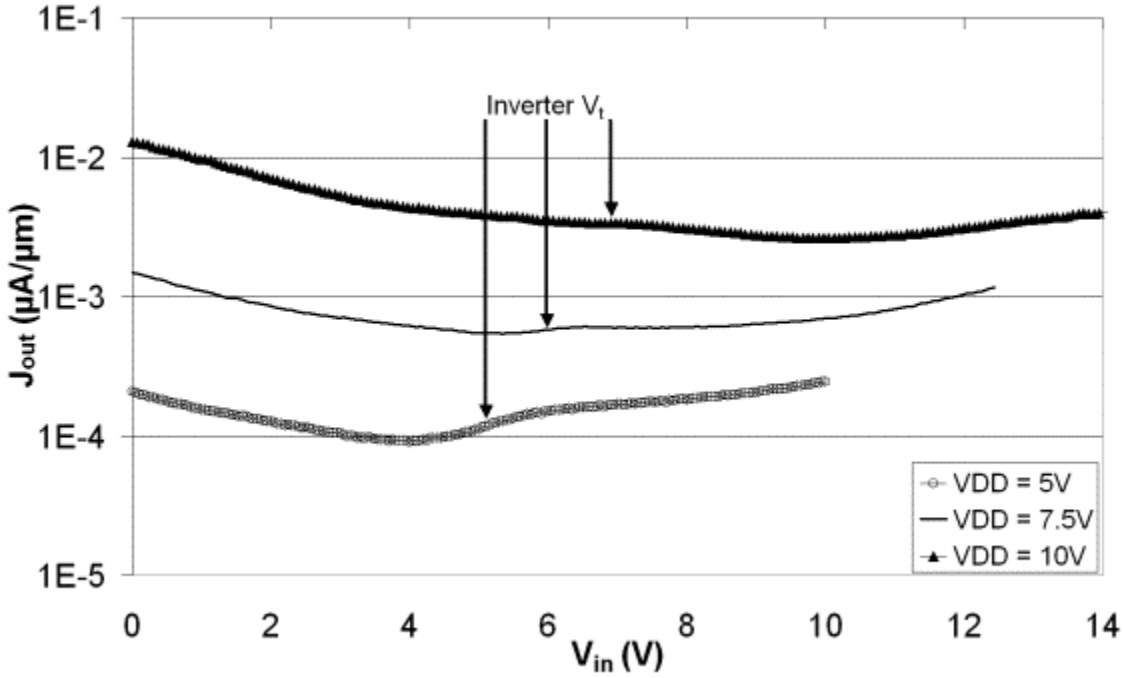

Fig. 9.5. Ouput current vs. V_{in} for the inverter in Fig. 9.3.

Fig. 9.5 shows the output current for the inverter in Fig. 9.3 as a function of V_{in}. Interestingly, this current does not modulate by much, and there exists no peak current corresponding to the inverter threshold (e.g., Fig. 7.18 in Chapter 7). As the current seems to increase for low and high V_{in}, it is quite possible that the ambipolar components observed in Fig. 9.4 at high V_{DS} play a significant role, although it is odd that the VTCs in Fig. 9.3 exhibit a near full swing with no indication of BBT (e.g., Fig. 7.18 in Chapter 7).

Fig. 9.6 shows the J_{DS} vs. V_{DS} for the NFET and PFET from Fig. 9.3. A very clear sub-linear region exists for both devices (although, predictably, more pronounced for the NFET), suggesting that a 5 min post-ITS anneal at 600 °C is not quite enough time to achieve "true" bulk switching characteristics, even for a boron-doped halo region. However, this is most likely attributable to the sub-optimal active region (Figs. 9.1 and 9.2), where the NiSi is very rough and may well be in a silicon-rich phase and/or is

highly discontinuous (i.e., mixed with pockets of silicon). This is supported by the 60-90 sec post-ITS anneals at 600 °C performed for discrete NFETs and PFETs in other work [1], [2], which achieved far greater performance than what is shown here.

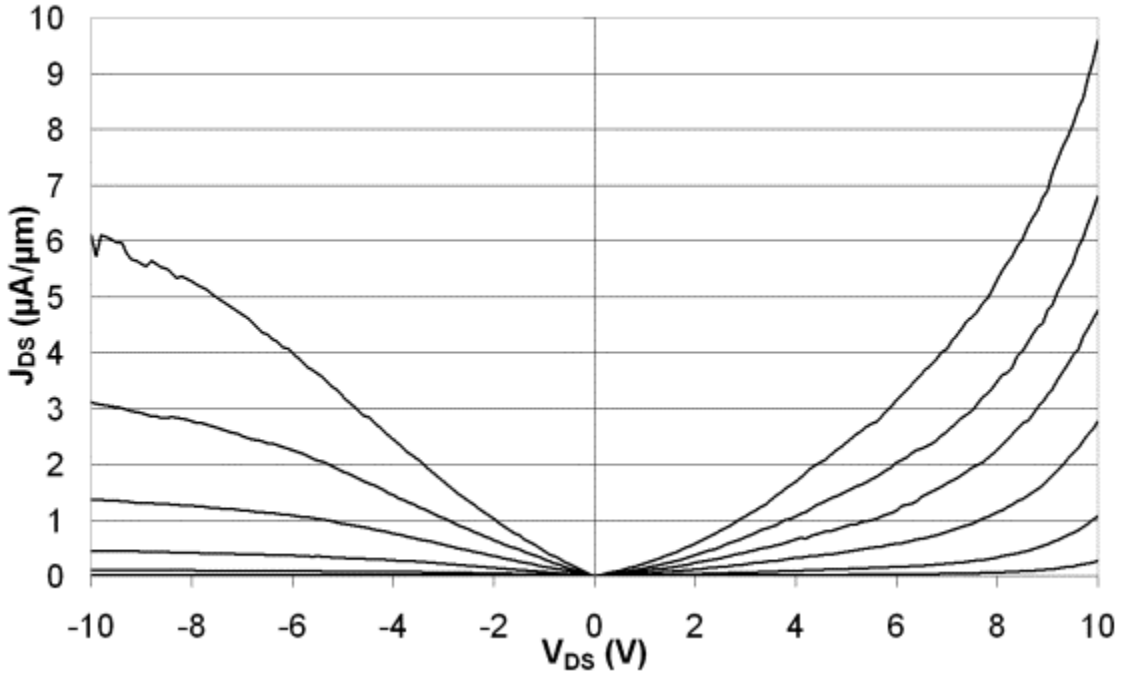

Fig. 9.6. J_{DS} vs. V_{DS} for the NFET and PFET from Fig. 9.3. $|V_{GS}|$ = 10-20 V in 2 V increments.

9.3 Suggestions for Future Studies

As alluded to previously, perhaps the most effective improvement that can be made to the process flow discussed in Section 9.1 is to change the active area anneal to something more likely to result in solid phase crystallization. This should greatly enhance the crystallinity, and hence performance, of these devices. Also, tailoring the phosphorus ITS to spread out within the entire silicide thickness by using either a higher implant energy and/or a lower body thickness, much like what was concluded for the SOI NFETs in Chapter 7, should greatly improve NFET performance. Beyond this, it would be very interesting to study the effect of time on device performance for a given post-ITS

anneal temperature, particularly for smaller times on the order of 1-2 min, to determine if any loss in drive current or increase in leakage current occurs compared to larger anneal times due to changes in W_{halo} and N_{halo}.

Chapter 9 References

[1] C.-P. Lin, Y.-H. Xiao, B.-Y. Tsui, "High-Performance Poly-Si TFTs Fabricated by Implant-to-Silicide Technique," *IEEE Elec. Dev. Lett.*, Vol. 26, no. 3, 2005, pp. 185-187.

[2] C.-F. Huang, B.-Y. Tsui, "Short-Channel Metal-Gate TFTs with Modified Schottky-Barrier Source/Drain," *IEEE Elec. Dev. Lett.*, Vol. 27, no. 1, 2006, pp. 43-45.

Chapter 10

Conclusions

10.1 Summary of Demonstrations and Findings

What has been presented in this work is, to the best of the author's knowledge, the first-ever full empirical demonstration of metallic source/drain (MSD) CMOS on both SOI and POI substrates, and a very clear demonstration of the relative ease with which CMOS is possible using the MSD approach. Particularly unique to this work is the utilization of FUSI source/drain regions also as inter-device isolation to facilitate circuit scaling without device scaling (and quite possibly without interconnect scaling) for at least one technology node. This alone is a valuable finding that may well ease some process constraints at very aggressive scales, as in the sub-100 nm regime, devices are scaled primarily to increase circuit density (and hence functionality) and not so much circuit speed. In addition, it has been shown that ITS processing has substantial potential for forming high quality source drain regions at low temperatures (~ 600 °C). This is very useful for both high performance CMOS with high-k gate dielectrics as well as TFT applications, where process temperatures may be limited by the thermal stability of the gate dielectric or underlying glass substrate, respectively.

By far, the most important finding of the presented work is that good Schottky CMOS is not Schottky CMOS, as gate modulation of a Schottky barrier is not nearly as efficient at modulating current flow as that of a conventional thermal barrier. Instead, this Schottky barrier must be minimized either by using rare earth metals, Group VI surface passivation, halo regions, or some combination thereof. Although the advantage

of using rare earth metals is the low intrinsic SBH to a particular carrier and the potential for completely implantless CMOS, the source/drain contact potential to the body region is the primary challenge to achieving an acceptable off state. This challenge is considerably reduced by utilizing halo regions, which can achieve low leakage with lightly doped or undoped body regions, while also reducing performance variability by essentially eliminating discrete dopant effects at small scales.

With regard to demonstrating CMOS in this study, the most important performance-affecting factor has been the limited diffusion of phosphorus within NiSi and the subsequent implications for NFET leakage and drive current. Empirically, this is in direct contradiction with the modeling results from Chapter 5, further demonstrating the limited insight achievable with said modeling. In any case, it seems quite clear that spreading out the as-implanted phosphorus profile throughout the entire silicide thickness should improve NFET performance in all aspects; this can be done by thinning the body region and/or increasing the implant energy. For the 100 nm body thickness used in this study, TRIM simulation suggests that increasing the implant energy from 34 keV to 80 keV will meet said requirement, albeit at the cost of a reduced peak halo concentration. Any effect of the reduced peak halo concentration on drive current, however, should be more than outweighed by the increased halo depth.

The smallest yielding MSD CMOS inverter in this study has a mask-defined gate length of 0.6 μm. Although the device performance is limited by the lack of a FUSI gate (thus increasing *EOT*) and the aforementioned NFET limitations, if the optimal process were used for the NFET and if a FUSI gate were achieved, it is very reasonable to suspect that, at $|V_{DS}| = |V_{GS} - V_{tlin}| = 5$ V, the PFET and NFET drive currents would be on the

order of 650 μA/μm and over 1 mA/μm, respectively. Although a minimal at best improvement over conventional CMOS at the same scale and voltage, the relative ease with which it can be achieved is quite startling.

10.2 Future Work

As with many endeavors with new technologies, the work presented here is hardly complete, and much remains to be done before one can claim a complete understanding and development of this device technology. There are three main areas of study within the presented body of work that deserve additional focus.

The first area is, quite naturally, modeling. Although it seems as if the developed Airy function model gives a more accurate solution to tunneling current than the WKB model, it is hardly done any justice by the simplistic energy band model used in Chapter 4. A self-consistent Poisson solution is the first step toward developing a more accurate SFET model, particularly if one wishes to model non-uniform body doping profiles (i.e., halo regions). In addition to this, for modeling that is relevant to nanoscale devices, quantum carrier confinement due both to thin body regions (UTBSOI or FinFETs) and high surface potentials must be accounted for, as well as channel mobility. Also, a better understanding of the nature of the halo regions (i.e., SIMS, SRP, etc. analysis) is necessary to appropriately model the effect of said regions on current injection through and over the Schottky barrier. This includes a consideration of bandgap narrowing for degenerately doped semiconductors, as well as a more accurate definition of the built-in voltage of the Schottky diode at such doping levels. As is becoming clear, most if not all of these model requirements do not permit closed form solutions, and so

the complexity of the modeling code will be considerably enhanced. Regardless, every single one of these ideas to explore has a significant dependence on the Schottky barrier height, and so it is critical to develop a more accurate model for Schottky barrier lowering as a function of the lateral field.

The second area of study is testing the theories developed about optimizing SOI and POI performance for the device structure presented. Most notably is determining if/how NFET performance is improved with a higher phosphorus implant energy into the NiSi, and whether post-ITS anneals of shorter time are actually feasible for POI substrates when the active region has been properly recrystallized.

The third, and possibly most important, area of study is comparing ITS vs. SIIS processing to form the halo regions (i.e., performing the halo implant after or before silicidation, respectively). Although the advantage of ITS is that the implant damage can be confined within the silicide, SIIS is possibly a better process to form smaller halo regions of higher concentration (while also reducing or eliminating the counterdoping effect at the V_{out} terminal mentioned in Chapters 6 and 7). The "expense" of SIIS is that post-implant anneals may need to be performed to mitigate defect-induced leakage, and this may compromise process compatibility with high-k gate dielectrics. Therefore, process splits with and without said anneal must be performed to compare diode ideality, reverse bias leakage, and ultimately FET performance.

Also, as the halo region becomes narrower, the tunnel barrier to minority carriers is reduced, thus resulting in almost the very same ambipolar leakage that the halo approach is meant to avoid in the first place. Higher halo concentrations increase the effective barrier height of this halo region to minority carriers, thus reducing tunneling

leakage, although there exists a limit as to the maximum achievable halo concentration (and this limit may well differ between ITS and SIIS due to the differing mechanisms of halo formation). Also, if the halo concentration is too high and the halo profile too abrupt, then band-to-band tunneling leakage will dominate over tunneling leakage through the halo region. Both modeling and experimental efforts must therefore be performed to determine the smallest halo width that results in acceptable leakage for a given halo concentration and how well this halo region can be controlled. Likewise, similar efforts must be performed to determine how device performance varies with halo width for a given halo concentration and profile of the tail region, as well as silicide material (e.g., how drive current may change in this architecture by using PtSi and $ErSi_2$). Ultimately, then, one must determine what "window" exists for these halo regions in both n-channel and p-channel nanoscale devices for different substrates and dopant species.

10.3 Closing Remarks

If one considers the challenges facing present and future CMOS scaling, it would seem that the MSD approach is inevitable at some point and in one form or another (i.e., ITS, SIIS, dual silicide, etc.). If anything, such a notion emphasizes the importance of studying MSD technology from all angles, and as soon as possible, considering the relatively small time between technology nodes. The advantages of the MSD architecture have been demonstrated or inferred to some capacity in the presented work, from both a device performance and a circuit density perspective. Although much remains to be done, this is not meant to undermine the value of the presented work, but rather to state that the best is yet to come.