Title
Advanced Source/Drain and Contact Design for Nanoscale CMOS

Author
Vega, Reinaldo

Publication Date
2010

Peer reviewed|Thesis/dissertation

Advanced Source/Drain and Contact Design for Nanoscale CMOS

by

Reinaldo Vega

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering-Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Tsu-Jae King Liu, Chair
Professor Chenming Hu
Professor Junqiao Wu

Spring 2010

Advanced Source/Drain and Contact Design for Nanoscale CMOS

by

Reinaldo Vega

Abstract

Advanced Source/Drain and Contact Design for Nanoscale CMOS

by

Reinaldo Vega

Doctor of Philosophy in Engineering – Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Tsu-Jae King Liu, Chair

The development of nanoscale MOSFETs has given rise to increased attention paid to the role of parasitic source/drain and contact resistance as a performance-limiting factor. Dopant-segregated Schottky (DSS) source/drain MOSFETs have become popular in recent years to address this series resistance issue, since DSS source/drain regions comprise primarily of metal or metal silicide. The small source/drain extension (SDE) regions extending from the metallic contact regions are an important design parameter in DSS MOSFETs, since their size and concentration affect contact resistance, series resistance, band-to-band tunneling (BTBT), SDE tunneling, and direct source-to-drain tunneling (DSDT) leakage. This work investigates key design issues surrounding DSS MOSFETs from both a modeling and experimental perspective, including the effect of SDE design on ambipolar leakage, the effect of random dopant fluctuation (RDF) on specific contact resistivity, 3D FinFET source/drain and contact design optimization, and experimental methods to achieve tuning of the SDE region.

It is found that DSS MOSFETs are appropriate for thin body high performance (HP) and low operating power (LOP) MOSFETs, but not low standby power (LSTP) MOSFETs, due to a trade-off between ambipolar leakage and contact resistance. It is also found that DSDT will not limit DSS MOSFET scalability, nor will RDF limit contact resistance scaling, at the end of the CMOS roadmap. Furthermore, it is found that SDE tunability in DSS MOSFETs is achievable in the real-world, for an implant-to-silicide (ITS) process, by employing fluorine implant prior to metal deposition and silicidation. This is found to open up the DSS process design space for the trade-off between SDE junction depth and contact resistance. $Si_{1-x}Ge_x$ process technology is also explored, and Ge melt processing is found to be a promising low-cost alternative to epitaxial $Si_{1-x}Ge_x$ growth for forming crystalline $Si_{1-x}Ge_x$ films.

Finally, a new device structure is proposed, wherein a bulk Tri-Gate MOSFET utilizes high-k trench isolation (HTI) to achieve enhanced control over short channel effects. This structure (the HTI MOSFET) is shown, through 3D TCAD modeling, to extend bulk

LSTP scalability to the end of the CMOS roadmap. In a direct performance comparison to FinFETs, the HTI MOSFET achieves competitive circuit delay.

_____

Professor Tsu-Jae King Liu

Committee Chair

To my teachers

Table of Contents

Acknowledgements

Many people struggle to find their path in life. For some, they cannot figure out how best to use their gifted mind. For others, they face the torture of one self-delusion after another, always thinking they have reached happiness but never actually feeling it. Some people stumble upon their passion, while for others it is force fed. I was one of those strange individuals who is just plain-old curious, and my pursuit of answers led me to my current position in life. I have been fortunate enough to meet along the way countless wonderful individuals who, knowingly or not, have shaped my life and my ambitions.

First on the list is my Ph.D. advisor, Prof. Tsu-Jae King Liu. I cannot say enough good things about her, and the praise she has endured from me and others is enough to fill a book. Few if any can match her technical prowess, and even fewer have the respect she has rightfully earned throughout her career. Her personal and professional advice will always remain with me, and her responsiveness, her professionalism, and her trust in me has been invaluable throughout my time here at UC Berkeley.

I am also grateful to Prof. Chenming Hu for his generosity and for the interesting discussions we have had. He has always managed to ask interesting questions about my research and, despite his remarkable knowledge, he is always open to new ideas. I would also like to thank Prof. Junqiao Wu for serving on my qualifying exam and dissertation committees, as well as Prof. Sayeef Salahuddin for serving on my qualifying exam committee, both of whom have provided valuable feedback.

I would like to thank all UC Berkeley Microlab staff for their hard work in keeping such a lively, borderline chaotic lab running smoothly. I would especially like to thank Bob Hamilton, Sia Parsa, Jimmy Chang, Danny Pestal, and Evan Stateler for their assistance with Tystar 19, GCAWS6, and Centura MXP. I would also like to thank Prof. Eugene Haller and his students – Swanee Shin and Christopher Liao – for their assistance with TEM analysis and understanding the Ge melt process. Furthermore, I would like to thank Salah Uddin, Raza Uddin, and Ning Chen, for their good company in the Microlab and many extended discussions/debates about cars and everything car-related.

Not all of my processing was performed in the Microlab, and so I would like to extend my appreciation to those vendors who helped me outsource some critical processes – TEM Analysis, QSpec, Core Systems, Evans Analytical Group (EAG), Martin Photomask, and UHV Sputtering. I would particularly like to thank Sanjay Patel at EAG for all his help in understanding SIMS analysis, and for the professionalism with which he has handled some very complex samples.

I am grateful to the people at the Semiconductor Research Corporation (SRC) and the MARCO/MSD Focus Center, and in particular Virginia Wiggins, Kim Wimberly, and Dr. Jeffrey Welser, for their assistance with various matters regarding my IBM/GRC Fellowship. I would also like to thank Ana Cargile, who first encouraged me to apply for this fellowship, as well as Prof. Karl Hirschman, Prof. Santosh Kurinec, Prof. James Moon, and Dr. Lisa Su, for supporting my GRC Fellowship and/or UC Berkeley applications. I also want to thank the GRC Fellows (Nicole Dilello, Sarah Bishop, Calvin King, Drew Forman, Crystal Kenney, Nathan Kupp, Dana Wheeler, Victoria Wang, and many others) and the large body of GRC students for their friendship and for stimulating discussions during SRC TECHCON and the MARCO/MSD grant reviews. The SRC community is full of wonderful people, and I am truly privileged to be part of it.

I also owe a debt to Dr. Michael Potter. He was a mentor of mine during my time at RIT and was relentless in his encouragement. He was the first person to motivate me to pursue a Ph.D. – "If you don't get a Ph.D., I'll kill you" – and his clarity of thought has always been an inspiration.

There are a ton of people at IBM who deserve thanks. I would first like to thank Dr. Edward Nowak and Dr. Gary Patton for agreeing to serve as my SRC Industrial Advisors. In particular, I have had many interactions with Dr. Nowak, and I very much enjoyed our numerous discussions and sharing my work with such a distinguished engineer. Dr. Patton opened my eyes to how rich a career one can have at IBM, and it was always a pleasure to speak with him. Also on the long list of current and former IBMers to whom I owe my gratitude are Dr. Melanie Sherony, Edward Maciejewski, Dr. George Hefferon, Kerry Leeburn, John Florkey, Kunal Vaed, Dr. William Ansley, Chris Schnabel, Robin Wanner, Dr. Noah Zamdmer, Lisa Hoysradt, Dr. Leland Chang, Dr. Wilfried Haensch, Dr. Christian Lavoie, Dr. Marwan Kahter, Dr. Tak Ning, Dr. Andres Bryant, Dr. Shreesh Narasimha, Dr. Alyssa Bonnoit, Kenji Yanagisawa, Dr. Lisa Edge, Dr. David Onsongo, Dr. Rama Divakaruni, Dr. Subramanian Iyer, and many others at Big Blue who made my various Co-Ops and visits enjoyable, informative, and challenging. I very much look forward to joining them again in the near future.

The list of fellow students at UC Berkeley who deserve my thanks is a very long one. I would first like to thank my senior students and industrial affiliates – Dr. Vidya Varadarajan, Dr. Sriram Balasubramanian, Dr. Donovan Lee, Dr. Andrew Carlson, Dr. Joanna Lai, Steve Volkman, Dr. Hiu Yung Wong, Dr. Kyoungsub Shin, Dr. Mohan Dunga, Dr. Carrie Low, Dr. Marie-Ange Eyoum, Dr. Pankaj Kalra, Dr. Alvaro Padilla, Dr. Noel Arellano, Dr. Vincent Pott, Dr. Alejandro de la Fuente Vornbrock, Dr. Chung-Hsun Lin, Dr. Wesley Chang, Dr. Tanvir Morshed, Dr. Vincent Pott, Yuri Masuoka, and Koichi Fukuda – for their wisdom, friendship, and technical advice. I would also like to thank my colleagues – Zach Jacobson, David Carlton, Darsen Lu, Xin Sun, Teymur Bakhishev, Anupuma Bowonder, Pratik Patel, Hei Kam, Changhwan Shin, Peter Matheu, Byron Ho, Nattapol Damrongplasit, Ryan Sochol, Adrienne Higa, Cheuk Chi Lo, Sung Hwan Kim, Min Hee Cho, Rhesa Nathaniel, Karl Skucha, John Gerling, Patrick Bennett, Sriramkumar Venugopalan, Shijing Yao, Jaeseok Jeon, Nuo Xu, Philip Chen, Wook Hyun Kwon, Lakshmi Jagannathan, Justin Valley, Amit Lakhani, Samuel Burden, Matthew Spencer, Alex Elium, Roger Chen, Amy Wu, Jodie Zhang, and many others – for their friendship, collaboration a variety of projects, and lively discussions/debates on a wide range of topics. They have made my time at UC Berkeley truly enjoyable.

I would like to give additional thanks to Alex Elium, Justin Valley, Amit Lakhani, Samuel Burden, Matthew Spencer, Zach Jacobson, Li-Wen Hung, Darsen Lu, and Roger Chen, for their help and enthusiasm regarding the outreach program that we all built together from scratch. I am very proud of this effort and its manifestation into something that reflects my vision on proper science and engineering outreach, and I am confident it will continue to grow in the right direction. I would also like to extend an enormous thanks to Mr. Matt McHugh of Berkeley High School, who has been very welcoming to us from the beginning and who is always excited to have our group visit. I will surely miss him and his classroom.

I owe a debt of gratitude to the undergraduate researchers who have worked in Prof. King's group during my time here – Lan Loi, Patricia Fong, Vincent Lee, Kevin Liu, Albert Lai, Helen Tran, and Alex Guo. I have enjoyed working with each of them and I wish them all the best.

Of course, I would not be here today were it not for my family. My parents' love, resilience, and never-ending commitment to do their very best at providing for me an opportunity to a future

they could only dream of, has set an example for me. I cannot thank them enough for their sacrifices, and I hope that I have made them proud. Lastly, but most certainly not least, I want to thank my wonderful girlfriend, Li-Wen Hung. My ambition to better myself and my research has brought me to Berkeley, and ultimately to her. She is the true prize of all my efforts and I love her with all my heart.

"We make our world significant by the courage of our questions and the depth of our answers."

- Carl Sagan

Chapter 1

Introduction

1.1 Moore's Law is a Byproduct

From 1965 onwards, a trend in integrated circuit (IC) manufacturing which later became known as Moore's Law [1] has more or less characterized the rate of growth of the semiconductor industry over time. Traditionally, this trend highlights the exponential growth of the component count on an IC; however, Moore's Law could equally apply to other computer metrics, such as storage density, computations per second, cost per function, etc.

However prescient the initial scaling predictions in [1] were (some would rightly argue the prediction was self-fulfilling), it would be disingenuous to motivate one's work in semiconductors with Moore's Law or even the ITRS Roadmap [2]. These are simply trends which, in the case of the former, illuminate the past and, in the case of the latter, attempt to place boundary conditions on the future. Most importantly, these trends describe precisely nothing about what these quantifiable advancements in technology have done for humanity, because ICs could be used for any number of things. Examples include ring oscillators (useless for computation), graphics processing units (GPU) and central processing units (CPU) for personal computers (PCs) and video game consoles, and distributed computing endeavors which aid in the search for gravity waves, extraterrestrial intelligence, and cures for various diseases [3]. Thus it would seem that simply quantifying the improvements in technology over time is not enough and that, instead, one must consider the improvements in functionality and/or productivity that have been achieved by the historical progression of IC technology. In other words, it is useful to understand what *motivates* scaling and it is equally important to understand that Moore's Law is a byproduct of this motivation. To this end, a historically and technically interesting example would be the scaling of video game technology. Having lived through this technological progression and experienced it firsthand as a consumer, the author resonates deeply with this example and so it is briefly discussed here.

The first commercially available home video game system was the Atari Pong, which was first produced in 1972. This "system" was actually only a single game, resembling a primitive form of tennis on a television screen. This was followed by the Atari 2600 in 1977, which popularized the concept of the gaming console – a single unit with exchangeable games, each stored on a read-only memory (ROM) cartridge. The Atari 2600 and its competitors at the time (*e.g.*, Mattel Intellivision, Emerson Arcadia 2001, etc.) are known as second-generation systems, having succeeded Pong. Third-generation systems, the most popular being the Nintendo Entertainment System (NES), released in 1985 in the U.S., improved upon the ROM formula. They did so by

not only increasing the color count to 52-256 (depending on the system) over the 16-color systems of second-generation consoles, but also through the introduction of accessory controllers such as running pads, infrared shooting guns, and glove controllers [4]. Some large-scale, quest-style games, such as Final Fantasy for the NES, introduced game save capability by integrating electrically erasable programmable read-only memory (EEPROM, also known as Flash memory) into the ROM cartridge. This feature may seem trivial today, but it was significant at the time, when resetting the console meant losing all progress in the game and starting over from the beginning.

It was arguably the success of the NES which set the stage for the multi-billion-dollar-per-year video game industry that would eventually materialize, since the NES amassed over its 9-year tenure a library of almost 800 games, the most popular of which (Super Mario Brothers) sold over 40 million copies worldwide [5]. To put this into perspective, the greatest-selling music album of all time – Michael Jackson's "Thriller" – has sold an estimated 110 million copies worldwide [6]. If sales volume is any indication of cultural impact, then surely the NES was a sign of things to come for video game culture.

By the late 1980's, the fourth generation of consoles began to emerge. These featured 16-bit processors and started with the Sega Genesis in 1988, followed by the Super Nintendo (SNES) in 1990. Other systems emerged as well, such as the SNK Neo Geo, but were a commercial failure due either to a small gaming library and/or a high selling price (*e.g.*, Neo Geo retailed for $650 in 1990, while SNES and Genesis retailed for $200). This is a recurring theme in the history of video games, in that the "console wars" are only played out among the two or three strongest manufacturers due to cross-platform software development costs, brand recognition, etc. Regardless, these fourth generation consoles continued to utilize ROM cartridges and so their only value-add to the consumer was higher color content (64-512 colors, depending on the system). This led to video games being memory-limited, and therefore content-limited on a fundamental level. Since ROM cartridges incorporate solid-state memory, read time is very short, which leads to near-instantaneous loading times for games. The caveat is that ROM cartridges suffer from lower memory density (fourth generation gaming cartridges had 4-6 MB of ROM) than emerging storage technologies at the time, namely, the compact disk (CD).

CDs are cheap, easy to manufacture, and can store 100x the memory as ROM cartridges (500-700 MB). However, in the late 1980s and early 1990s, CD drives were first-generation and ran at what is now called 1x, meaning a maximum read speed of 153.6 kB/s. Thus, loading an entire game into system memory would take at least 26 sec, an intolerable wait time for impatient gamers. The memory bottleneck induced by ROM cartridges left console manufacturers with no choice, however, and in 1991 the Sega CD was released as an add-on to the Sega Genesis console. Although not a market success, others followed (NEC TurboGrafx-CD, Philips CDi, Panasonic 3DO, NEC TurboDuo, and SNK Neo Geo CD) until eventually the Sony PlayStation (PS1), released in the U.S. in 1995, marked the beginning of the fifth generation era of video game consoles.

The fifth generation of consoles was defined chiefly by optical storage, 32-bit processors, and 3D rendering. Also, removable Flash memory cartridges made game saves commonplace. In fairness, however, 3D rendering began on the SNES with the release of Star Fox in 1993. Although manufactured on a ROM cartridge, it featured the "Super FX" chip, which was a reduced instruction set computing (RISC) math co-processor designed as a graphics accelerator chip (one of the very first graphics accelerators, if not the first, for mainstream 3D rendering). Although the 3D renderings had a low polygon count (due to the small amount of memory on the

ROM cartridge), the game was tremendously popular and marked the beginning of 3D gaming and graphics acceleration which the fifth-generation gaming consoles would take to an entirely new level.

Not surprisingly, 3D rendering and optical storage were complementary in fifth generation consoles. The high storage capacity of CDs over ROM cartridges permitted a substantial increase in gaming content, since the high polygon count for detailed 3D renderings and the images that must be mapped to the polygon surfaces consume substantial memory (high quality audio and video recordings supplemented the 3D environments to increase gaming content and consume even more memory). Fifth-generation consoles, such as the PS1 and Sega Saturn (Nintendo continued with the ROM formula for their fifth generation console, the Nintendo 64), utilized 1x CD drives, much like the fourth generation console add-ons. Load times were long, but the considerable increase in gaming content over fourth generation consoles made it worth the wait. Thus, the market success of the PS1 far overshadowed its hardware deficiency of a slow optical drive. By the time the sixth generation of consoles began to be released in the late 1990s (Sega Dreamcast, 1999), gaming content had become too rich for CDs in the same way that ROM cartridges were becoming outdated in the early 1990s. Yet another shift in storage technology was imminent – the digital video disc (DVD), which offers ~10x more storage than the CD (4.7-8.5 GB).

The most successful sixth generation consoles (Sony PlayStation 2, or PS2, released in 2000 and the Microsoft Xbox, released in 2001) utilized DVD storage and faster optical drives to reduce load time. At this point, the gaming content permitted by DVD storage had become so rich that CPU and GPU designs were becoming very complex. The "Emotion Engine" used in the PS2 was fabricated on a 250 nm process and contained 10.5 million transistors (compared to the Intel Pentium II which, while also fabricated on a 250 nm process, contained only 7.5 million transistors). This was an important time in video game history, because gaming consoles had become so advanced that their hardware performance was, in some ways, on par with or in excess of mainstream personal computers (PCs). Recognizing this trend, Microsoft saw fit to make the Xbox nothing more than a cheap PC (literally, it used an Intel Pentium III CPU and Nvidia GeForce 3 GPU, practically off the shelf), in order to minimize hardware development costs. The seventh-generation consoles to follow (Microsoft Xbox 360, released in 2005, the Sony PlayStation 3, or PS3, released in 2006, and the Nintendo Wii, released in 2006) would continue this trend. The concomitant increase in computational power in these seventh generation consoles was accompanied by yet another upgrade in optical storage technology – Blu-Ray (PS3) and high-definition DVD (HD-DVD, for the Xbox 360) – each offering another ~10x increase in storage over DVDs. It is worth taking a brief pause here, to reflect on how seventh generation game storage media hold up to 50-100 GB of memory (dual-layer Blu-Ray), compared to 4 kB in the Atari 2600. Over six generations of gaming consoles, the information content (or at least the potential content) within each game increased by a factor of over 10 million and computational power had to scale accordingly, just to keep up.

Although both the Xbox 360 and PS3 are strong competitors in the seventh generation "console wars," it is arguably the PS3 which has wider-reaching significance. Development on the PS3 began in 2000, around the same time the PS2 was released, in recognition of the monumental task of designing a successful seventh generation gaming console. The Cell microprocessor, the CPU used in the PS3, was developed at a cost of U.S. $400 million with a broad range of multimedia and other computationally expensive applications in mind [7]. At the turn of the century, when mainstream single-core CPUs were struggling to reach GHz clock

speeds, when computational tricks such as pipelining and branch prediction were heavily utilized in next-generation CPU architectures, the Cell offered a vision of the future – a whopping eight processor cores all running in parallel, with a ninth core as the controller. This was arguably the first mainstream implementation of parallel processing, and on a level of parallelization that, in the year 2010, remains unmatched (the AMD Opteron 24xx series, with six cores, comes closest). Thus, the PS3 is computationally significant not just for its gaming power, but in its utility in high performance computing for non-gaming applications. For the first time ever, people have begun to take seriously the idea of using video game consoles for distributed computing and cost-effective supercomputer clusters [8], [9]. The sheer scale and potential of the PS3 gives a glimpse of the influence that video game technology and scaling has, and will continue to have, on humanity's pursuit of the ultimate calculator.

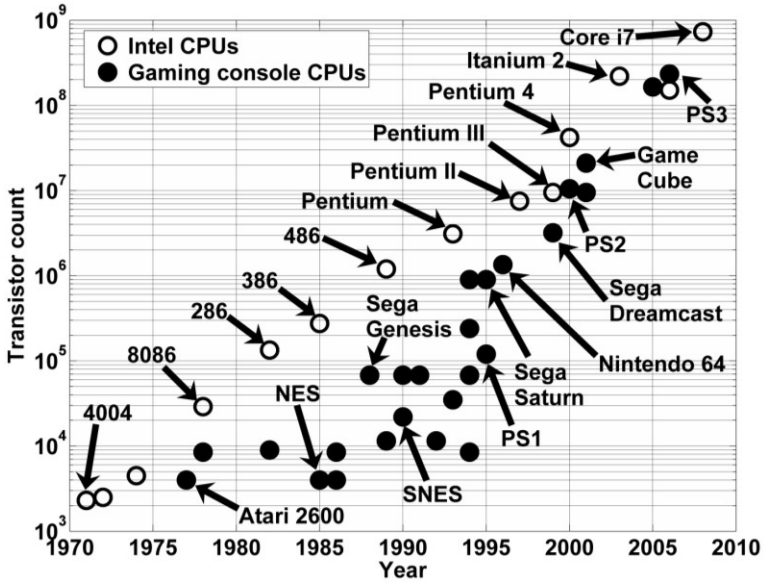

Fig. 1.1. Transistor count vs. time, comparing Intel CPUs to video game console CPUs.

Thus, the scaling of video game technology was driven by complementary scaling in memory storage and computational power (*i.e.*, CMOS scaling), all in the name of providing a more immersive, more realistic entertainment experience. As Fig. 1.1 shows, from ~1990 onwards, the scaling rate in transistor count for video game console CPUs was higher than that of Intel's mainstream CPUs. While the transistor count in Intel's CPUs increased by ~ 2 decades per 10 years, for gaming consoles this rate was 2.5-3 decades per 10 years (from 1990 onwards). In fairness, this rate of scaling had more to do with console manufacturers shifting from using pre-existing microprocessors to developing their own microprocessors (*e.g.*, the MOS 650X series, Motorola 68000, and Zilog Z80 microprocessors were commonly used in gaming consoles for several generations, up to the mid-1990s) than it had to do with an above average rate of CMOS scaling. Nevertheless, current generation consoles are so complex that their transistor count is on par with that of modern desktop PCs and substantial investments by the video game industry (*e.g.*, U.S. $400 million by Sony [7] and U.S. $1 billion by Nintendo [10]) to develop these consoles shows that video games are now a significant driver for CMOS scaling.

1.2 Motivation for Alternative Device Structures

Having established at least one motivation for CMOS scaling, it is now worthwhile to consider some challenges the CMOS industry faces moving forward and to discuss how to address them. At the most basic level, the goal of CMOS scaling is to make a more efficient switch. Since the energy required to switch a transistor on or off is equal to $0.5*C_G*V_{DD}^2$, where C_G is the gate capacitance and V_{DD} is the power supply voltage, CMOS scaling has traditionally involved reducing both the gate length L_G (which proportionally reduces C_G) and V_{DD}. However, V_{DD} scaling is not very aggressive and has been delayed (Fig. 1.2). This is because MOSFETs are thermal devices, meaning the ideal subthreshold swing (SS) is limited to (kT/q)*ln(10) or 60 mV/dec at 300 K. As V_{DD} is reduced for the same off-state current I_{OFF} (and therefore the same threshold voltage V_T), the on-state current I_{ON} drops due to reduced gate overdrive ($V_{GS} - V_T$). Likewise, maintaining I_{ON} at reduced V_{DD} means reducing V_T, thereby increasing I_{OFF}, which leads to higher standby power dissipation. Either way, V_{DD} scaling is necessary in order to reduce power density on an IC, since increased power density leads to heat, which reduces device performance (due to reduced mobility) and further increases standby power consumption (due to increased thermal leakage).

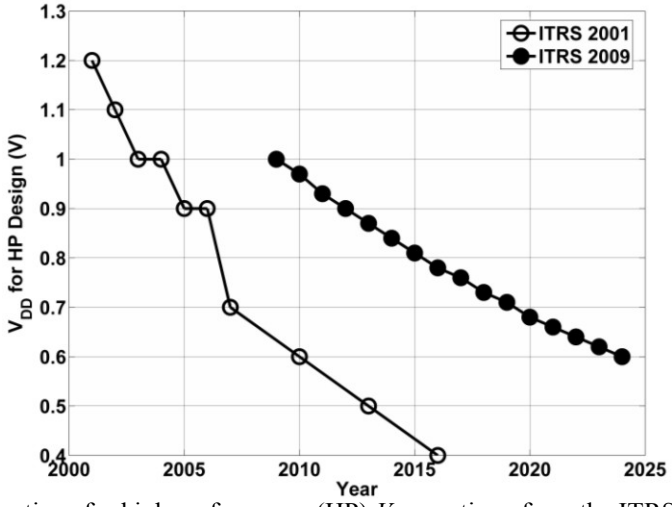

Fig. 1.2. ITRS specifications for high performance (HP) V_{DD} vs. time, from the ITRS 2001 and 2009 roadmaps.

Thus, to the device designer, the present and future goal of CMOS scaling is keep I_{ON} as high as possible while scaling V_{DD}. This means reducing or eliminating resistance in the source/drain and channel regions, but in a manner that does not sacrifice MOSFET switching integrity, regardless of the device structure in question (*e.g.*, planar bulk, planar SOI, multi-gate MOSFETs, etc.). Reducing L_G will reduce channel resistance due to the shorter path length between the source and drain regions, with further channel resistance reduction possible through strain engineering [11] or alternative substrate materials [12]. However, the source/drain extension junction depth $X_{j,SDE}$ must scale with L_G in order to maintain switching integrity. For the same source/drain doping profile, the source/drain sheet resistance $R_{s,SDE}$ will increase [13] according to Equations (1.1) and (1.2), where q is the electron charge, $\mu(N,x)$ is the doping- and position-dependent mobility at 300 K for n-type Si [14], and $N(x)$ is the position-dependent doping concentration.

$$R_{s,SDE} = \left[\int_0^{X_{j,SDE}} q\mu(N,x)N(x)dx \right]^{-1} \tag{1.1}$$

$$\mu(N,x) = 88 + \frac{1250}{1 + \frac{0.88 * N(x)}{1.26 * 10^{17}}} \tag{1.2}$$

This increase in $R_{s,SDE}$ can be compensated by increasing the source/drain doping concentration, but this requires higher temperature processing which leads to increased dopant diffusion, thus increasing $X_{j,SDE}$. Several doping and annealing schemes, such as plasma doping [15], [16], gas phase doping [17], [18], and pulsed laser anneal (PLA) [19] have been proposed in order to achieve high source/drain doping concentration while minimizing dopant diffusion. To illustrate how heavily-doped the SDE regions must be for future technology nodes, a fully-depleted SOI (FDSOI) case is considered, whereby the body thickness t_{body} is small enough such that the SDE doping distribution is uniform (*i.e.*, *N(x)* is constant and $t_{body} = X_{j,SDE}$). Using the relationship in Equation (1.3) for characteristic length λ (where ε_{gd} is the gate dielectric constant and EOT is the equivalent oxide thickness), and noting that $L_G \geq 5\lambda$ for good control over short channel effects [20], one can plot theoretical $R_{s,SDE}$ vs. L_G for different doping concentrations and compare these curves to the 2009 ITRS Front End Processing specifications [2]. This is shown in Fig. 1.3. For EOT = 1 nm, shown in Fig. 1.3(a), t_{body} must be small to satisfy $L_G \geq 5\lambda$, which means the SDE doping must be very high, on the order of 3-4x10^{20} cm^{-3}. This is close to the electrically active solid solubility limit of n-type dopants at 1000 ºC [13]. Shrinking EOT to 0.5 nm, as in Fig. 1.3(b), reduces this doping requirement to ~2x10^{20} cm^{-3}, which is more accessible but still technically challenging.

$$\lambda = \frac{L_G}{5} = \sqrt{\frac{\varepsilon_{body}}{\varepsilon_{gd}} EOT * t_{body}} \tag{1.3}$$

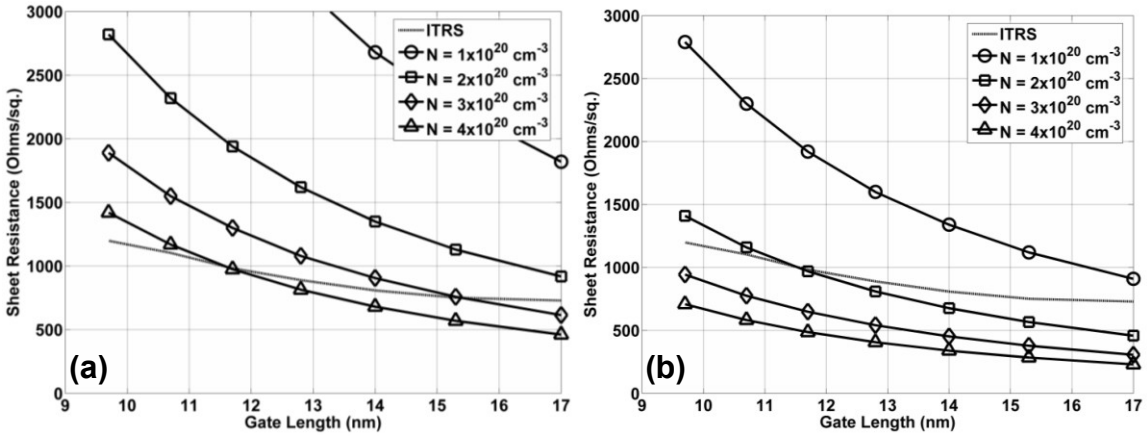

Fig. 1.3. Theoretical SDE sheet resistance vs. L_G for different doping levels, compared to ITRS 2009 FDSOI specifications, for (a) EOT = 1 nm and (b) EOT = 0.5 nm.

No matter how heavily the source/drain region is doped, though, it is not as conductive as metal. In comparison to the data in Fig. 1.3, the ITRS target silicide sheet resistance is 8.7 Ω/sq.

at L_G = 17 nm, increasing to 14.9 Ω/sq. at L_G = 9.7 nm. R_s < 20 Ω/sq. is easily achieved for NiSi at ~20-25 nm thickness [21], [22] and the ITRS target values are less than 1.5 % of the target $R_{s,SDE}$ values, which raises the question: why not make the source/drain regions out of metal silicide, instead of doped silicon?

The idea of using metal silicide as the source/drain region, known as metallic source/drain (MSD) CMOS, is nothing new. The first MSD MOSFET was proposed in 1968, using PtSi source/drain regions to demonstrate PMOSFET behavior [23]. This device was a subset of MSD MOSFETs known as Schottky barrier (SB) MOSFETs, wherein the silicide forms a SB contact to the channel, which is modulated by the gate voltage. Subsequent efforts emerged in the mid-1990s [24] and gained popularity due to the simple process flow compared to that for conventional doped source/drain MOSFETs, with SB MOSFETs demonstrated with L_G as low as 10 nm [25]. As will be discussed in subsequent chapters, though, the performance of a SB MOSFET is limited by the SB height (SBH) at the silicide/silicon interface, such that the resulting increase in contact resistance more than offsets the low silicide sheet resistance for a net reduction in I_{ON}.

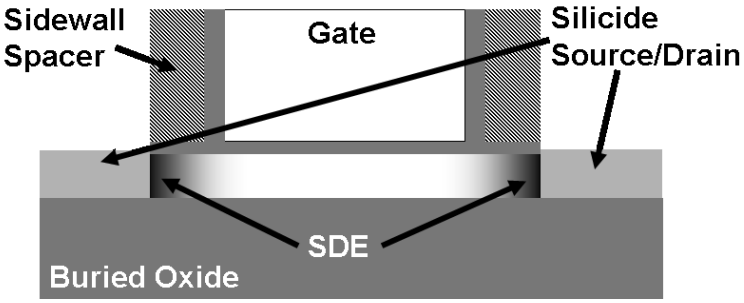

Fig. 1.4. Schematic cross-section of a DSS MOSFET on SOI.

Another subset of MSD MOSFETs, known as dopant-segregated Schottky or DSS MOSFETs, was proposed in the early 2000s [26] and is a more promising approach toward MSD CMOS. In this type of device, a shallow, heavily-doped SDE region lies adjacent to the silicide region (Fig. 1.4), in order to reduce the contact SBH [26]. This type of structure is arguably nothing more than a conventional MOSFET, since it contains a doped SDE region; however, with the majority of the volume of the source/drain region constituting metal or metal silicide, the device is technically a MSD MOSFET. Semantics aside, though, there are many questions that remain about DSS MOSFET design and whether DSS is an appropriate source/drain technology for any or all power/performance specifications (*i.e.*, HP, low standby power or LSTP, low operating power or LOP). These questions include, but are not limited to, addressing ambipolar leakage mechanisms, the effect of random dopant fluctuation (RDF) on specific contact resistivity (ρ_c), single silicide CMOS vs. dual silicide CMOS, and process integration schemes that permit tuning of $X_{j,SDE}$ in DSS MOSFETs.

1.3 Dissertation Objectives and Outline

The ultimate objective of this dissertation is two-fold: 1) to determine, through modeling and experiment, whether NiSi is suitable for single-silicide CMOS to the end of the CMOS roadmap and 2) to explore DSS and other MOSFET designs at and near the end of the CMOS roadmap.

In Chapter 2, DSS MOSFETs are compared to raised source/drain (RSD) MOSFETs for LSTP and HP design, using 2D and 3D TCAD modeling. It is shown that fundamentally different source/drain architectures are optimal for different power/performance specifications. For LSTP design, RSD MOSFETs are more appropriate than DSS, since they can be tuned to achieve a larger leakage floor (I_{min}) design space and higher I_{ON}. For HP design, a version of DSS MOSFETs called recessed strap (RS) DSS is most suitable, since it combines the DC and AC performance merits of both conventional DSS and RSD MOSFETs to result in a device structure with a universal performance advantage over other source/drain designs.

Chapter 3 explores DSS MOSFET design in the sub-10 nm L_G regime (using 2D TCAD), where direct source-to-drain tunneling (DSDT) is expected to be significant. It is shown that L_G is not an appropriate design metric in this regime and must be co-optimized with the other device geometries ($X_{j,SDE}$, sidewall spacer length, etc.), and that it is possible to design against DSDT to the point where it does not contribute to I_{OFF}. At this point, dual high-k/low-k sidewall spacers are introduced as a critical element to enabling MOSFET scaling into this regime, as it provides an alternative to body thickness and/or $X_{j,SDE}$ scaling.

In Chapter 4, an analytical model is developed to describe the effect of RDF on ρ_c. This model is calibrated against experimental data on NiSi and PtSi contacts to n- and p-type Si and includes all SB lowering effects – image force, interface dipole, and bandgap narrowing. It is shown that ρ_c variation due to RDF drops as the doping concentration at the contact interface is increased and that, with a doping level of ~2×10^{20} cm^{-3}, NiSi contacts to Si can achieve $\rho_c < 10^{-8}$ Ω-cm^2 with low variation, even for contact areas at the end of the CMOS roadmap (20-30 nm^2).

Chapter 5 introduces a new device structure, the high-k trench isolation (HTI) bulk Tri-Gate MOSFET. This structure utilizes high-k dielectrics as the shallow trench isolation (STI) material (or as the STI liner) to amplify the reverse narrow width effect, in essence creating a device which performs competitively with the FinFET without the process complexities associated with FinFETs. It is shown that this HTI MOSFET can extend bulk LSTP scalability far beyond ITRS projections, all the way to the end of the CMOS roadmap.

Chapter 6 explores implant-to-silicide (ITS) process technology as an avenue for fabricating DSS MOSFETs. DSS junction formation is shown to be directly related to silicide thermal stability and, as a consequence, exerting control over the silicide thermal stability allows for modulation of $X_{j,SDE}$. DSS diodes, NMOSFETs, and PMOSFETs are fabricated and characterized, with the experimental results in agreement with DSS junction formation theory. This lends experimental evidence to the notion that NiSi, properly utilized, can be suitable for ultimate junction scaling toward the end of the CMOS roadmap.

In Chapter 7, $Si_{1-x}Ge_x$ process technology is explored. Low-temperature LPCVD of n- and p-type $Si_{1-x}Ge_x$ is investigated over a range of Ge contents, and it is found that n- and p-type $Si_{1-x}Ge_x$ deposition requires much more than simply changing the dopant carrier gas. Additionally, methods to form crystalline $Si_{1-x}Ge_x$ from LPCVD $Si_{1-x}Ge_x$ and Ge films are investigated. Solid phase epitaxial regrowth (SPER) is limited by the high parasitic O content in the LPCVD films, while Ge melt processing suffers from a Ge spiking effect when RTA is used to form the melt. A slower, quasi-steady-state furnace melt is shown to eliminate the Ge spiking effect.

Chapter 8 summarizes the contributions of this dissertation and discusses future research prospects for the topics presented in this study.

1.4 References

[1] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, pp. 114-117, 1965.

[2] International Technology Roadmap for Semiconductors (ITRS). [Online]. Available: http://public.itrs.net

[3] D. P. Anderson, "BOINC: A System for Public-Resource Computing and Storage," *Proceedings of the 5th IEEE/ACM International Workshop on Grid Computing*, pp. 4-10, 2004.

[4] List of Nintendo Entertainment System accessories. [Online]. Available: http://en.wikipedia.org/wiki/List_of_Nintendo_Entertainment_System_accessories

[5] Guiness World Records Gamer's Edition. [Online]. Available: http://gamers.guinnessworldrecords.com/records/nintendo.aspx

[6] K. Anderson, "Michael Jackson's *Thriller* Set to Become Top-Selling Album of All Time." [Online]. Available: http://www.mtv.com/news/articles/1616537/20090720/jackson_michael.jhtml

[7] J. A. Kahle, M. N. Day, H. P. Hofstee, C. R. Johns, T. R. Maeurer, D. Shippy, "Introduction to the Cell multiprocessor," *IBM J. Res. Dev.*, vol. 49, no. 4/5, pp. 589-604, July/Sept. 2005.

[8] J. Wilson, M. Dai, E. Kakupovic, S. Watson, F. Meng, "Supercomputing with Toys: Harnessing the Power of Nvidia 8800GTX and Playstation 3 for Bioinformatics Problems," *Comput. Syst. Bioinformatics Conf.*, pp. 387-390, 2007.

[9] J. Kurzak, A. Buttari, P. Luszczek, J. Dongarra, "The PlayStation 3 for High Performance Scientific Computing," *Computing in Science and Engineering*, vol. 10, no. 3, pp. 84-87, 2008.

[10] R. Wilson, "IBM grabs next Nintendo system win," *EE Times*, May 1999. [Online]. Available: http://www.eetimes.com/story/OEG19990512S0025

[11] G. Eneman, E. Simoen, P. Verheyen, K. De Meyer, "Gate Influence on the Layout Sensitivity of $Si_{1-x}Ge_x$ S/D and $Si_{1-y}C_y$ S/D Transistors Including an Analytical Model," *IEEE Trans. Elec. Dev.*, vol. 55, no. 10, pp. 2703-2711, Oct. 2008.

[12] D. A. Antoniadis, A. Khakifirooz, "MOSFET Performance Scaling: Limitations and Future Options," *IEDM Tech. Dig.*, pp. 253-256, 2008.

[13] R. C. Jaeger, "Introduction to Microelectronic Fabrication Volume V," *Addison Wesley Publishing Co.*, pp. 58-72, 1993.

[14] R. S. Muller, T. I. Kamins, M. Chan, "Device Electronics for Integrated Circuits," *John Wiley & Sons, Inc.*, p. 33, 2003.

[15] K. Kobayashi, K. Okuyama, H. Sunami, "Plasma doping induced damages associated with source/drain formation in three-dimensional beam-channel MOS transistor," *Microelectronic Engineering*, vol. 84, pp. 1631-1634, 2007.

[16] B. Mizuno, Y.Sasaki, "Aiming for The Best Matching between Ultra-Shallow Doping and Milli- to Femto-Second Activation," *IEEE Advanced Thermal Processing of Semiconductors*, pp. 1-10, 2007.

[17] C. M. Ransom, T. N. Jackson, J. F. DeGelormo, C. Zeller, D. E. Kotecki, C. Graimann, D. K. Sadana, J. Benedict, "Shallow n+ Junctions in Silicon by Arsenic Gas-Phase Doping," *J. Electrochem. Soc.*, vol. 141, no. 5, pp. 1378-1381, May 1994.

[18] J. C. Ho, R. Yerushalmi, Z. A. Jacobson, Z. Fan, R. L. Alley, A. Javey, "Controlled nanoscale doping of semiconductors via molecular monolayers," *Nature Materials*, vol. 7, pp. 62-67, Jan. 2008.

[19] Y. F. Chong, K. L. Pey, A. T. S. Wee, A. See, L. Chan, Y. F. Lu, W. D. Song, L. H. Chua, "Annealing of ultrashallow p^+/n junction by 248 nm excimer laser and rapid thermal processing with different preamorphization depths," *Appl. Phys. Lett.*, vol. 76, no. 22, pp. 3197-3199, May 2000.

[20] 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 Sub-Micrometer MOSFETs," *IEEE Trans. Elec. Dev.*, vol. 40, no. 1, pp. 86-95, Jan. 1993.

[21] A. Lauwers, J. A. Kittl, M. Van Dal, O. Chamirian, R. Lindsay, M. de Potter, C. Demeurisse, C. Vrancken, K. Maex, X. Pages, K. Van der Jeugd, V. Kuznetsov, E. Granneman, "Low temperature spike anneal for Ni-silicide formation," *Microelectronic Engineering*, vol. 76, pp. 303-310, 2004.

[22] M. Tsuchiaki, K. Ohuchi, A. Nishiyama, "Suppression of Thermally Induced Leakage of NiSi-Silicided Shallow Junctions by Pre-Silicide Fluorine Implantation," *Jpn. J. Appl. Phys.*, vol. 44, no. 4A, pp. 1673-1681, 2005.

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

[24] J. R. Tucker, C. Wang, P. S. Carney, "Silicon field-effect transistor based on quantum tunneling," *Appl. Phys. Lett.*, vol. 65, no. 5, pp. 618-620, Aug. 1994.

[25] M. Jang, Y. Kim, M. Jun, C. Choi, T. Kim, B. Park, S. Lee, "Schottky Barrier MOSFETs with High Current Drivability for Nano-regime Applications," *Journal of Semiconductor Technology and Science*, vol. 6, no. 1, pp. 10-15, Mar. 2006.

[26] 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," *IEDM Tech. Dig.*, 2004, pp. 168-169.

Chapter 2

FinFET Source/Drain Design Optimization

2.1 Introduction

In recent years, there has been a growing interest in metallic source/drain (MSD) MOSFETs [1]–[12], wherein the potential exists for reduced source/drain resistance (R_{SD}) and increased immunity to process variation due to the elimination of source/drain (S/D) dopants. Early studies focused on Schottky source/drain architectures on bulk substrates [4], [5], for which the gate directly modulates the source-body Schottky barrier (SB). However, leakage current was high due to subsurface source-body leakage and ambipolar tunneling injection through the drain-body SB. Also, the on state current (I_{ON}) is limited by the SB height (SBH), and so a need for a lower on-state SBH and/or better SB modulation as well as lower leakage became apparent. Subsequent modeling work highlighted the efficacy of ultra-thin body (UTB) regions and/or reduced equivalent oxide thickness (*EOT*) for increasing I_{ON} in Schottky-S/D MOSFETs [13]. The UTB structure also reduces leakage due to the elimination of a sub-surface region, as shown in [3]. However, in that study, the leakage floor (I_{min}) was still too high for low standby power (LSTP) applications [14], and I_{ON} was still SB limited, even with near-band-edge silicides.

Given that Fermi-level pinning in practice limits the SBH to values that are too high to achieve high I_{ON}, regardless of the metal or silicide work function [1-6], the next step in the evolution of MSD MOSFET design was to passivate or otherwise modify the metal-semiconductor (M-S) interface to reduce the SBH to electrons or holes for NMOS or PMOS devices, respectively. Some have demonstrated, through modeling or experiment, the use of thin interfacial layers [15], [16] or Group VI valence-mending adsorbates [17-21] with promising results. However, the interfacial layer approach has limited integration potential due to its process complexity, and also limits I_{ON} due to the tunnel barrier imposed by the interfacial layer. The use of Group VI elements is tantamount to using dopants to form a source/drain extension (SDE) at the M-S interface, since both approaches have the same effect on the interface dipole [21], [22]. Using dopants at the interface, however, has an added benefit of further reducing the SBH through image field barrier lowering if the SDE region is long enough and/or doped highly enough so that it is not fully depleted.

An additional advantage of using dopants at the M-S interface, as opposed to an interfacial layer or Group VI passivation, is that the width (as well as the height) of the tunneling barrier to minority carriers at the drain (e.g., holes for NMOS) is

increased, which reduces ambipolar leakage (normally SB tunneling, but now SDE tunneling) and therefore I_{min}. This effect has been achieved with both implant-to/through-silicide (ITS) [7], [8] and dopant pile-up during silicidation [9-11], [23-25], although the results have been mixed and the dominant leakage mechanism has been difficult to determine with certainty due to the lack of accurate information about the lateral dopant profiles in these devices. In other words, while in all cases a notable increase in I_{ON} was achieved, the measured I_{min} values and ambipolar behavior have varied, which is a clear indication of the sensitivity of I_{min} to the lateral dopant profile, gate over/underlap, trap density within the SDE region, *etc*. An understanding of the leakage behavior (*i.e.*, SB tunneling, SDE tunneling, and band-to-band tunneling) is necessary to guide the LSTP design of MSD MOSFETs with SDE regions (*i.e.*, dopant-segregated Schottky or DSS MOSFETs).

For high performance (HP) design, however, limitations imposed by leakage constraints are less likely to apply, and the interest lies primarily in the ability of SDE dopants to reduce the SBH for improved I_{ON}. Thus for HP design, the tradeoff is between reduced source/drain resistance but reduced contact area for DSS MOSFETs vs. increased source/drain resistance but increased contact area for raised source/drain (RSD) MOSFETs. Although published work has consistently reported superior performance for DSS MOSFETs as compared with doped-source/drain MOSFETs [26]-[30], the actual performance gain was usually marginal and the cause was either unclear (*i.e.,* the devices were not properly optimized) or trivial (*i.e.*, more aggressive silicide formation). Nevertheless, such work suggests that fabrication is a simpler endeavor for DSS MOSFETs than for doped-source/drain MOSFETs, and that design optimization for high performance (HP) applications may be more straightforward for DSS *vs.* doped-source/drain devices. For thin-body transistors such as FinFETs, though, the DSS structure may be more difficult to fabricate due to the need for very thin metal deposition to prevent voiding or excessive lateral silicidation [31]-[33]. Assuming that advancements in metal deposition technology will address this issue, the remaining question is how DSS *vs*. RSD FinFETs compare in terms of AC performance, considering their three-dimensional (3D) structural details which affect parasitic capacitance and hence delay. Thus, the goal of this study is to investigate, through 2D and 3D TCAD using Sentaurus Device [34], whether DSS FinFETs are competitive with RSD FinFETs at aggressive scales, for both LSTP and HP design.

2.2 LSTP Design Study

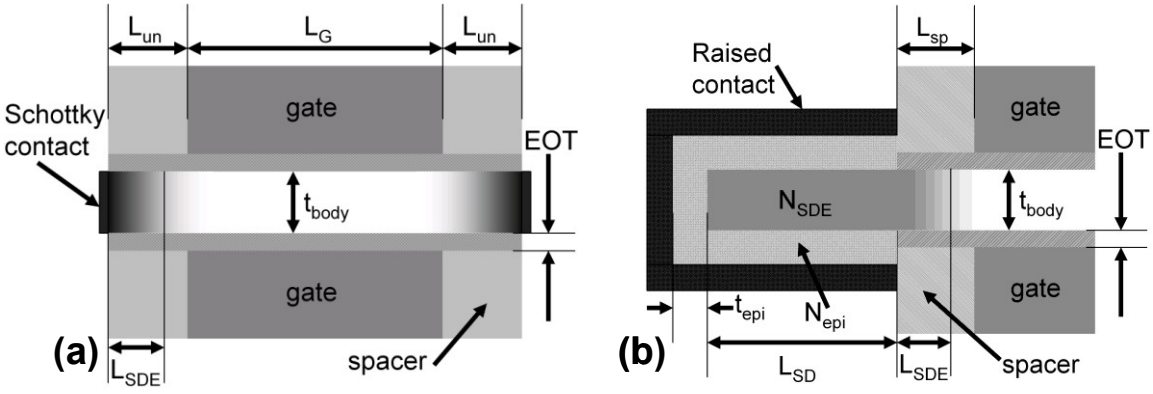

Fig. 2.1. Schematic cross-sections of the (a) DSS and (b) RSD structures modeled in this study.

The symmetric double gate (SDG) NMOSFET is considered here, with the DSS structure shown in Fig. 2.1(a) and the RSD structure shown in Fig. 2.1(b). The body thickness t_{body} = 10 nm unless otherwise noted, EOT = 1 nm (gate leakage is ignored here), and the p-type body doping is $1x10^{15}$ cm^{-3}. A metal gate is used, ideal V_t tuning through gate work function engineering is assumed, the sidewall spacer material is silicon nitride, and V_{DD} = 1 V.

Two MSD materials are studied – NiSi and $ErSi_{1.7}$. For NiSi, the electron SBH is 0.65 eV [17] and the hole SBH is 0.47 eV, while for $ErSi_{1.7}$, the electron SBH is 0.3 eV and the hole SBH is 0.82 eV. Empirically, the SBH values for $ErSi_{1.7}$ are very sensitive to interface states/defect density [3], [35]-[40], so here a value of 0.3 eV for the electron SBH is chosen, considering the variation in reported data. The SDE region is n-type and has a doping profile with peak concentration N_{SDE} at the M-S interface, and decaying as a Gaussian function away from this interface. The SDE length L_{SDE} is defined as the distance from the gate-sidewall spacer edge (corresponding to the location of the M-S junction) to where the SDE concentration drops to $1x10^{18}$ cm^{-3} (the cut-off between degenerate and non-degenerate doping), and is varied from 3-10 nm. This gives junction abruptness values between ~1.67 and 10 nm/dec over the N_{SDE} range studied. For the RSD structure in Fig. 2.1(b), the SDE region extends into the silicon that is not overlapped by the gate-sidewall spacer (with a length L_{SD} = 3*L_G), and this part of the SDE region has constant doping of N_{SDE}. The epitaxial RSD region in Fig. 2.1(b) has a thickness t_{epi} (varied from 2-10 nm) and a constant doping profile with concentration N_{epi}.

The maximum gate underlap L_{un} (negative overlap) to the M-S junction considered here is 10 nm. The notation L_{un} is used in Fig. 2.1(a), while in Fig. 2.1(b) the notation L_{sp} is used to represent the gate-sidewall spacer length. Both notations represent the length of the spacer, although in practice L_{un} for a DSS MOSFET can be smaller than the spacer length, due to lateral silicidation. In the RSD structure, however, the same SDE profile would require a smaller spacer since the silicide is located too far away to effectively pile-up dopants laterally in the SDE region; hence the different notations.

2.2.1 Simulation Setup

1-D non-local tunneling models were used for all tunneling calculations. Band-gap narrowing (BGN) as a function of doping concentration was included. The SDE regions are treated as ideal (*i.e.*, with zero trap density, which would otherwise be due to incomplete post-implant annealing) and the SBs are treated as abrupt changes in potential (*i.e.*, no rounding of the potential profile near the M-S interface).

SBL due to dipole and image field effects [41], [42] is excluded. Since both the dipole and image field SBL calculations have an electric field dependence [41], [42], and the electric field at the interface changes as the SBH is reduced due to the presence of dopants at the M-S interface (*i.e.*, due to dipole-induced SBL), a self-consistent solution to the potential at and near the M-S interface is required. Unfortunately, such a self-consistent SBL calculation is not implemented currently in the TCAD software. Even if such an SBL model were available, it must be excluded due to an accuracy tradeoff in calculating the on-state and off-state currents. This tradeoff is described as follows.

To calculate the SB tunneling current, a Poisson-Schrödinger solution was initially used with electron and hole effective tunneling masses of 0.19*m_0 and 0.16*m_0, respectively (assuming

that electron transport is dominated by the Δ2 subband and hole transport is dominated by the light hole subband). The effective Richardson's constants for electrons and holes were set to 112 and 32 A/cm^2K^2, respectively. To calculate band-to-band tunneling (BTBT) current simultaneously with SB/SDE tunneling, convergence issues prevent the use of the Schrödinger solution with the BTBT model, and so the use of a simpler (but less accurate) model for SB/SDE tunneling – the Wentzel-Kramers-Brillouin (WKB) approximation – is required. Additionally, the BTBT model is implemented with a two-band dispersion relation, since the energy of the BTBT carriers lies far from the band edges. However, the two-band dispersion relation tends to inflate SB/SDE tunneling, which is already inflated by the WKB model [43]-[45]. This overestimation of SB/SDE tunneling relative to the Schrödinger solution is compensated here by increasing the effective tunneling masses for electrons and holes in the WKB model. This results in a WKB-to-Schrödinger modeling fit in the low-field/wide-barrier regime in which the SDE doping is low enough such that BTBT does not determine I_{min} (*i.e.*, SB or SDE tunneling determines I_{min}). Here, the electron and hole effective masses were tuned to $0.42*m_0$ and $0.36*m_0$, respectively, for the WKB model. The resulting fit is shown in Fig. 2.2 for both NiSi and $ErSi_{1.7}$ MSD regions and a relatively light SDE doping of $1x10^{19}$ cm^{-3}, which clearly shows the Schrödinger solution and the "tuned" WKB model giving the same result for I_{min}.

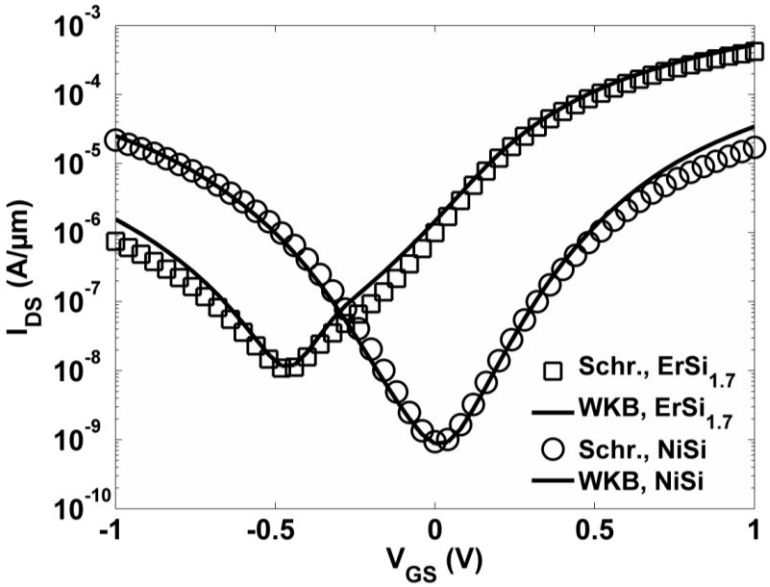

Fig. 2.2. I_{DS} vs. V_{GS} for SDG DSS MOSFETs with NiSi or $ErSi_{1.7}$ MSD regions and V_{DS} = 1 V. L_G = 15 nm, L_{SDE} = 5 nm, L_{un} = 2 nm, t_{body} = 10 nm, EOT = 1 nm, and N_{SDE} = $1x10^{19}$ cm^{-3}. The gate workfunction is set to 4.1 eV. For the Schrödinger solution, the electron and hole effective masses are $0.19*m_0$ and $0.16*m_0$, respectively. For the WKB solution with the two-band dispersion relation, the electron and hole effective masses are $0.42*m_0$ and $0.36*m_0$, respectively.

Despite the match in I_{min} results achieved, the WKB and Schrödinger tunneling solutions always diverge in the on-state for any effective mass value, as Fig. 2.2 shows, because in this region the "wide barrier" assumption [44] inherent in the WKB model is no longer valid (i.e., electron tunneling through a thin SB is modeled in the on-state, whereas the off-state involves hole tunneling through a wider tunneling barrier). This reduces the accuracy of the WKB model in predicting I_{ON}, but since SBL is not modeled, one can simplistically treat this divergence as the effect of SBL on I_{ON}. The ratio of I_{ON} values at V_{GS} = 1 V between the WKB and

Schrödinger models is 1.25x for $ErSi_{1.7}$ and 2x for NiSi. This is qualitatively consistent with the trend of SBL and tunneling, each contributing less to I_{ON} for lower SBH. Therefore, I_{min} and trends in I_{ON} can be modeled accurately with both BTBT current and SB/SDE tunneling in LSTP devices. Admittedly, a deficiency of this approach is that it ignores the effect of SBL to reduce ambipolar leakage in the off state. For example, in the case of an NMOS device, as the electron SBH is lowered due to the presence of n-type dopants, the hole SBH is raised by the same amount; this changes the shape of the tunneling barrier to holes at the drain (widening the barrier at lower electron energies) as Fig. 2.3 shows. Since this effect is excluded here, the simulated dependence of I_{min} on N_{SDE} is shifted slightly to higher doping values than if SBL were included; however, the trends observed and the conclusions reached do not change but are instead reinforced due to the contact-limited performance of DSS MOSFETs, as will be shown.

It is important to stress that, although I_{ON} is calculated here using the tuned WKB tunneling model, by no means does this approach stipulate that tunneling is the dominant current component in the on state. Tunneling current is strongly dependent on barrier height [43] and effective tunneling mass [46]. Thermal injection over the Schottky barrier most likely dominates I_{ON}, at least for the case where the doped SDE lowers the SBH. However, since SBL is not modeled in the LSTP study, the increase in thermal current that would otherwise occur is captured by the aforementioned trend in the WKB model in the on-state.

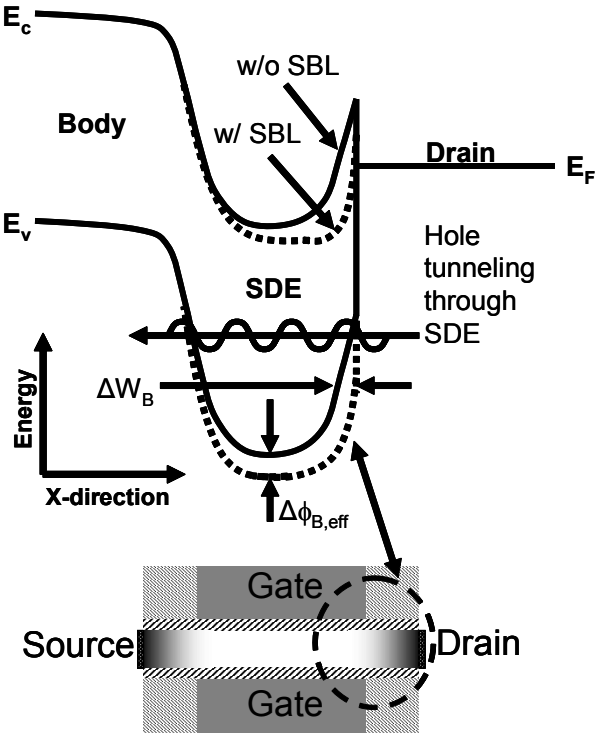

Fig. 2.3. Energy band diagram in the drain region of a DSS MOSFET with PD SDE regions. L_{un} and L_{SDE} are arbitrary. The dotted lines illustrate the change in the band diagram when SBL is considered, and how this affects the tunneling barrier height $\phi_{B,eff}$ and width W_B for holes at low electron energies in the drain region.

To model carrier transport, a conventional drift-diffusion transport model with dopant- and field-dependent mobility is used. In [47], an increase in injection velocity beyond the thermal velocity for DSS MOSFETs was reported and attributed to SB transport at the source junction; however, this is likely not the case. The physical explanation provided in [47] for the relative

performance improvement with DSS MOSFETs is that the SDE region induces a high electric field at the SB at the source (and drain). When electrons tunnel through the source-side SB, they are accelerated by the high field at the SB to velocities beyond the thermal velocity (*i.e.*, velocity overshoot), thus improving I_{ON}. However, this explanation neglects the effect of the SDE regions on SBL, which reduces the SBH and therefore the electric field at the source. It also excludes the effect of silicide formation on tensile stress in the channel, which becomes significant for NiSi with source-to-drain silicide spacings of less than 100 nm and is stronger for more rounded silicide profiles, where the tensile stress adjacent to the silicide can be on the order of 100 MPa or greater [48], [49]. Given that the silicide formation in the DSS structure in [47] is more aggressive than that for the conventional MOSFET, and that the silicide profile is more rounded, it is entirely possible that the measured increase in injection velocity has more to do with a stress-induced increase in thermal velocity than with carrier acceleration at the source-side SB. Thus the DSS vs. conventional MOSFET behavior in Fig. 5(b) in Ref. [47] can be explained by an increase in channel stress in the DSS structure with device scaling. Also, the reduced source-to-drain silicide spacing in the DSS structure in [47] results in an increase in the lateral field for the same channel voltage V_{ch} (the drain-source voltage minus the voltage drop across the parasitic source/drain resistance), which confounds the effect of stress to further increase drive current. This is not to state that the field at the source-side SB plays no role whatsoever in carrier acceleration, but rather to state that its role has yet to be quantified accurately, and that likely its role is very small due to the effect of the SDE region on reducing the SBH to very small values (as will be shown in Chapter 4). In the presented modeling study, the effect of silicide-induced stress is excluded, since in principle it provides no performance improvement that could not be realized in RSD MOSFETs through other methods (*e.g.*, stress liners, source/drain stressors, etc.).

2.2.2 Effect of N_{SDE} on Leakage

It has been stated that a fully-depleted (FD) SDE region in MSD MOSFETs maintains the "merits" of a Schottky junction [50]. However, Schottky junctions have no merits as S/D junctions in nanometer-scale MOSFETs because, as discussed previously, they reduce I_{ON} and increase subthreshold swing (*SS*), while increasing I_{min} due to tunneling injection at the drain. A non-fully-depleted (partially depleted or PD) SDE region is therefore required, because this increases the SB/SDE tunneling barrier width and height at the drain, thus reducing leakage (*i.e.*, SB tunneling is the leakage mechanism for FD SDE regions, where the depletion region width $W_{D,SB}$ extending from the metal junction exceeds the SDE junction depth $x_{j,SDE}$, while SDE tunneling is the leakage mechanism for PD SDE regions, where $W_{D,SB} < x_{j,SDE}$). Also, the higher doping required to achieve a PD SDE region is advantageous because it further shrinks the SB, which increases I_{ON}. To determine whether the SDE region is FD or PD, one can plot *SS* or I_{min} *vs*. N_{SDE}, as shown in Fig. 2.4 and Fig. 2.5, respectively.

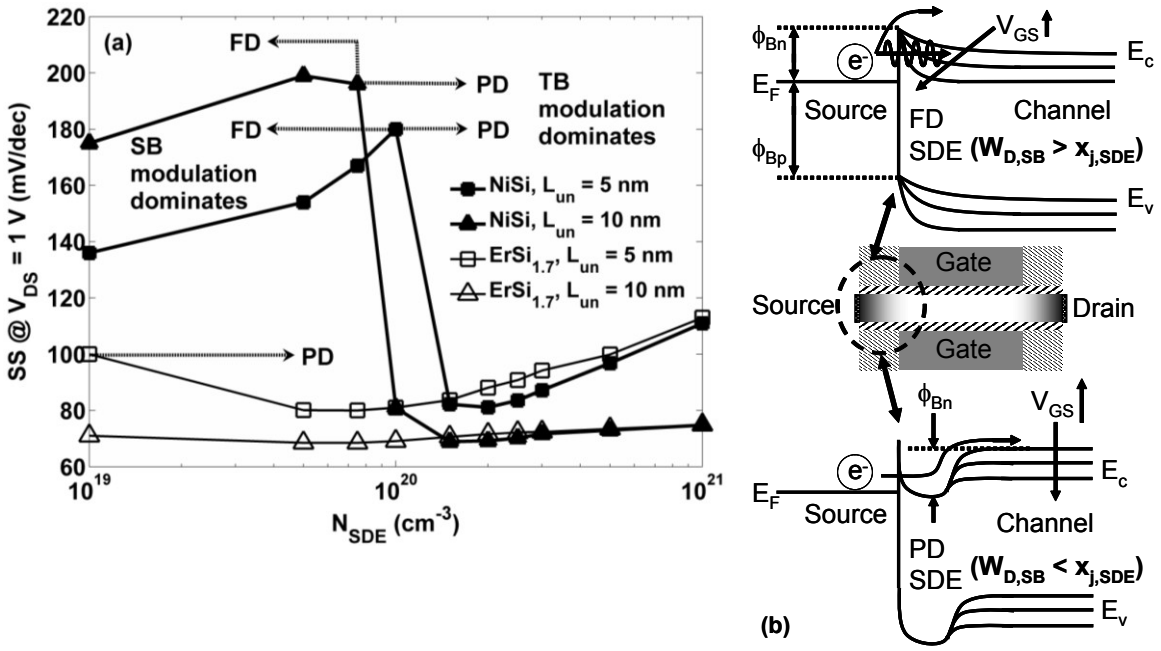

Fig. 2.4. *SS* vs. N_{SDE} for SDG FETs with NiSi or $ErSi_{1.7}$ MSD regions. L_G = 15 nm, t_{body} = 10 nm, L_{SDE} = 3 nm, *EOT* = 1 nm, and V_{DS} = 1 V. As N_{SDE} is increased, the subthreshold behavior moves from SB modulation (FD SDE) to thermal barrier (TB) modulation (PD SDE). (b) Energy band diagrams illustrating current injection at the source for FD and PD SDE regions, and how the source-to-body barriers change with applied gate bias V_{GS}.

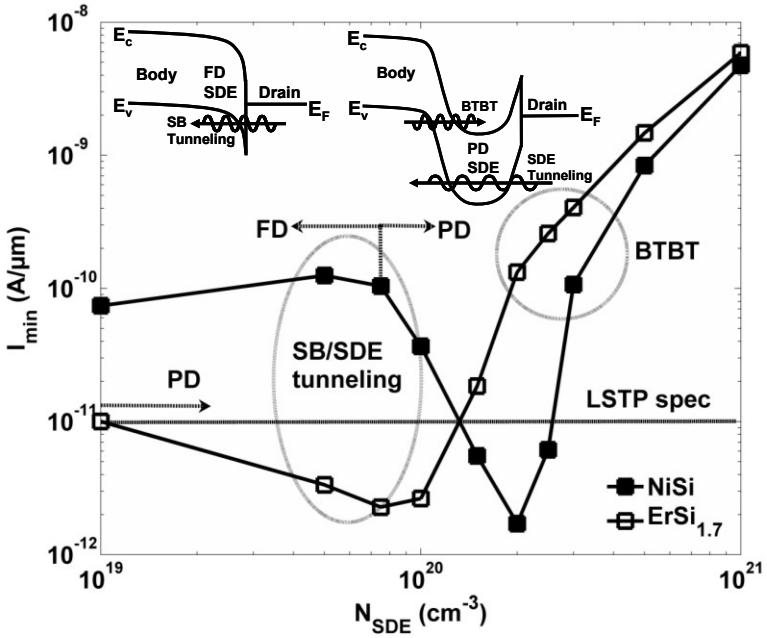

Fig. 2.5. I_{min} vs. N_{SDE} for SDG FETs with NiSi or $ErSi_{1.7}$ MSD regions. L_{un} = 10 nm and the remaining device parameters are the same as for Fig. 5. Energy band diagrams at the top of the figure illustrate the various leakage mechanisms.

As Fig. 2.4(a) shows, *SS* for the NiSi source/drain regions is high for low N_{SDE}, and increases with L_{un}. This indicates that the SDE regions are FD, and that the source-body SB limits current flow, as the top of Fig. 2.4(b) shows. As L_{un} increases, the SB is placed farther away from the gate edge, and so the gate fringing field has less control over modulating the SB. As N_{SDE} is increased, *SS* increases so long as the SDE regions remain FD, due to an increase in charge sharing between the gate and FD SDE regions. When N_{SDE} is increased such that the SDE region shifts from FD to PD, *SS* drops substantially because it is now determined by conventional thermal barrier (TB) modulation rather than SB modulation, as Fig. 2.4(a) and the bottom of Fig. 2.4(b) show. Note that in this case *SS* decreases with increasing L_{un} due to improved control of short-channel effects (SCE) and that the transition from FD to PD takes place at higher N_{SDE} when L_{un} is reduced due to the gate playing an increased role in depleting the SDE region. Further increasing N_{SDE} increases *SS* again due to worsening SCE, just as in a conventional MOSFET. For $ErSi_{1.7}$ MSD regions, the SBH is lower, so the depletion region extending from the M-S junction into the SDE region is smaller [51]. This is why *SS* for the $ErSi_{1.7}$ case in Fig. 2.4 is determined by TB modulation (as evidenced by the reduction in *SS* with increasing L_{un}) over the entire range of N_{SDE} studied.

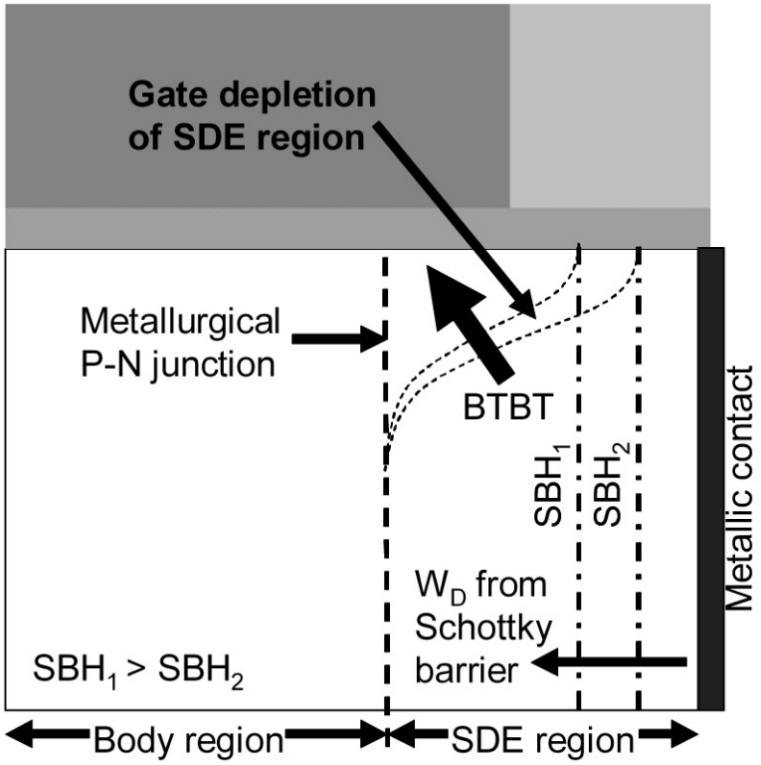

Fig. 2.6. Illustration of BTBT in devices with PD SDE regions.

In Fig. 2.5, I_{min} is determined by SB/SDE tunneling at the drain, for low values of N_{SDE}. In this regime, I_{min} follows the same trend as *SS* for both NiSi and $ErSi_{1.7}$, where it increases with N_{SDE} for NiSi (due to SCE combined with the smaller tunnel barrier for the FD SDE region) and decreases for $ErSi_{1.7}$ (due to the larger tunneling barrier for the PD SDE region). When N_{SDE} is further increased to result in the FD-to-PD transition for NiSi (7.5×10^{19} cm^{-3} for L_{un} = 10 nm), I_{min} drops sharply with increasing N_{SDE}, since both the tunneling barrier width and height at the drain increase. The width increases due to the size of the quasineutral part of the PD SDE

region, while the height increases due to the dopant-induced energy-band bending (*i.e.*, diode built-in voltage V_{bi} when the system is in equilibrium) that adds to the hole SBH to result in a taller barrier. However, as N_{SDE} is further increased, I_{min} increases again due to BTBT. The lowest I_{min} is therefore a balance between SDE tunneling and BTBT, resulting in a small N_{SDE} window for LSTP design.

It is interesting to note that, for the same N_{SDE}, BTBT current is higher for $ErSi_{1.7}$ than for NiSi, as Fig. 2.5 shows. The higher BTBT for the $ErSi_{1.7}$ case is attributable to the lower SBH to majority carriers, which reduces the extent of the SDE depletion for a given N_{SDE}. The less depleted the SDE region is, the greater the cross-sectional surface area for BTBT to take place, due to the gate-induced depletion of the SDE region near the surface. This is illustrated in Fig. 2.6. Although the higher BTBT component limits N_{SDE} to a lower value for the case of $ErSi_{1.7}$ for LSTP design, a net gain in I_{ON} is still achieved over the case of NiSi due to the lower SBH (shown later).

2.2.3 Effect of L_{SDE} on Leakage

An additional approach to reduce SDE tunneling at the drain is to increase L_{SDE}, which increases the tunneling barrier width. This is why deep source/drain regions are required for bulk DSS MOSFETs [9], [52], [53], because otherwise drain-to-bulk SDE tunneling leakage increases I_{min}. SDG MOSFETs do not have deep source/drain regions; however, increasing the size of the SDE region has the same effect. (This can also be achieved electrostatically for conventional SB MOSFETs, as in [54].) This is shown in Fig. 2.7(a) for NiSi S/D regions and Fig. 2.7(b) for $ErSi_{1.7}$ S/D regions. For low N_{SDE}, tunneling through the SDE determines I_{min}, so that it can be reduced by increasing L_{SDE} (*e.g.*, by ~ 1 order of magnitude when L_{SDE} is increased from 3 nm to 7 nm, in the NiSi MSD case). As N_{SDE} increases, increasing L_{SDE} (*i.e.,* broadening the SDE doping profile) eventually results in the gate overlapping the SDE-body junction enough to allow BTBT to dominate over SDE tunneling, so that I_{min} increases with L_{SDE}.

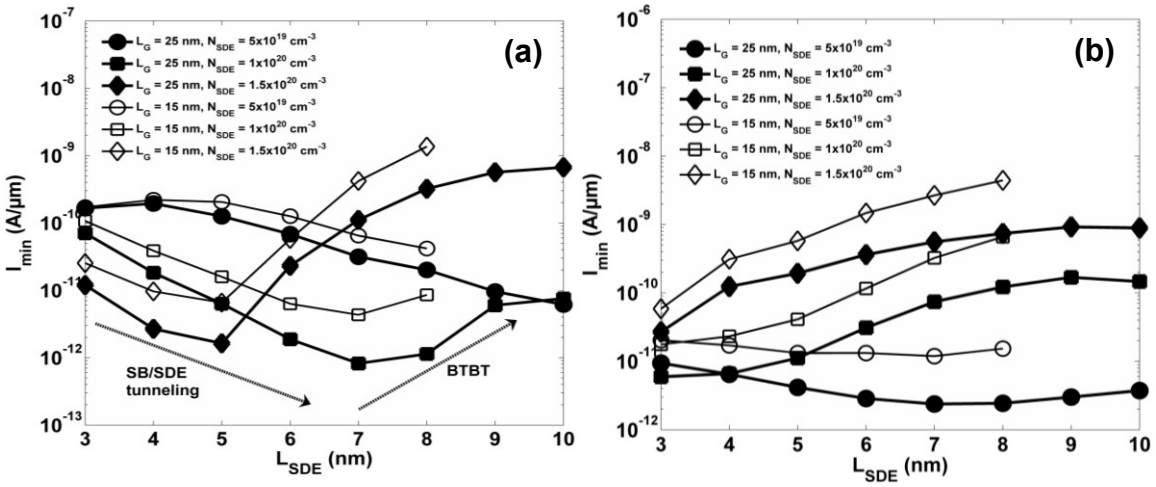

Fig. 2.7. I_{min} vs. L_{SDE} for SDG FETs with (a) NiSi and (b) $ErSi_{1.7}$ MSD regions, and N_{SDE} = 5x10^{19}- 1.5x10^{20} cm^{-3}. EOT = 1 nm, t_{body} = 10 nm, L_{un} = 8 nm, and V_{DS} = 1 V.

The reduction in SDE tunneling with increasing L_{SDE} is not quite as large for $ErSi_{1.7}$ as it is for NiSi and is due to the larger BTBT component, as discussed previously regarding Figs. 2.5 and 2.6. Note that, to keep the same I_{min} and with all other device parameters constant, scaling L_G requires an increase in L_{SDE} in the regime where SDE tunneling dominates (due to the increased lateral field for the same V_{DD}), while the opposite is the case in the regime where BTBT dominates (to reduce the gate overlap of the SDE region). It is clear from Fig. 2.7 that I_{min} is much more sensitive to N_{SDE} than L_{SDE}, which intuitively follows from the WKB approximation, which indicates the tunneling probability dependence on barrier height ϕ_B (determined by N_{SDE}, as mentioned earlier) to be $exp(-\phi_B^{3/2})$, and the dependence on W_B (determined by L_{SDE}) to be $exp(-W_B)$.

2.2.4 Effect of N_{SDE} and L_{SDE} on I_{ON}

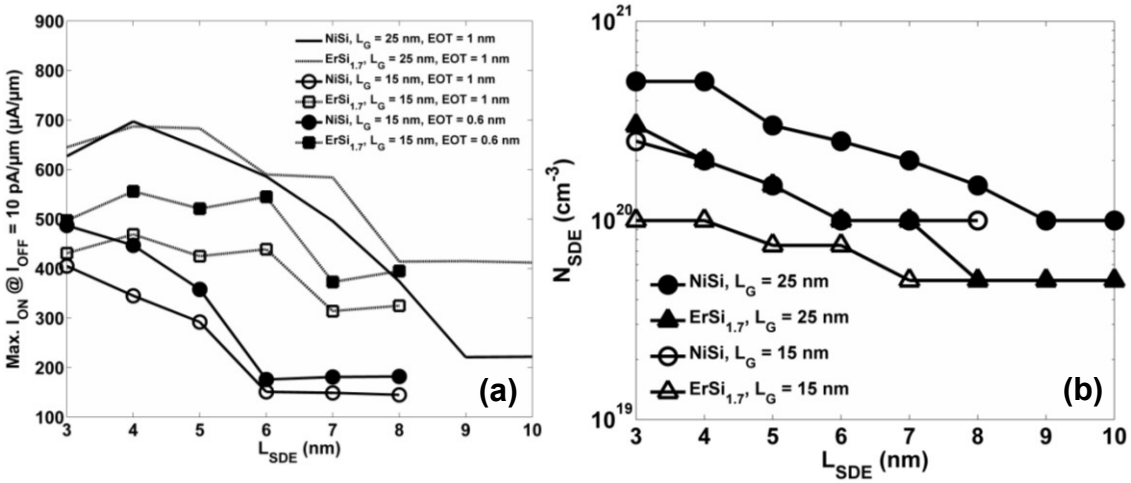

Fig. 2.8. (a) maximum LSTP I_{ON} vs. L_{SDE} and (b) corresponding optimal N_{SDE} vs. L_{SDE} for SDG FETs with NiSi or $ErSi_{1.7}$ MSD regions, for various EOT and L_G. In all cases, t_{body} = 10 nm and V_{DD} = 1 V. The optimal N_{SDE} values are the same for EOT = 1 nm and 0.6 nm.

Fig. 2.8(a) shows the best-case I_{ON} vs. L_{SDE} for constant I_{OFF} = 10 pA/μm, where the best-case I_{ON} is determined for each L_{SDE} by finding the optimal combination of N_{SDE} (Fig. 2.8(b)) and L_{un}. In most cases, although not shown in any of the figures, the optimal L_{un} = 10 nm, but occasionally drops to as low as 7 nm, primarily due to the N_{SDE} resolution in this modeling study. Thus L_{un} can be treated as constant, which suggests that the optimal I_{ON} is determined chiefly by L_{SDE} and N_{SDE}. As L_{SDE} is reduced, the concomitant reduction in SCE permits a higher N_{SDE} (Fig. 2.8(b)) and, to a lesser extent, lower L_{un}, both of which improve I_{ON}. Thus, DSS MOSFETs utilizing near-midgap silicides such as NiSi can be competitive with devices utilizing near-band-edge silicides such as $ErSi_{1.7}$, provided L_{SDE} is small enough and N_{SDE} is high enough.

As L_G is scaled, for constant L_{SDE}, the reduction in optimal N_{SDE} due to SCE (Fig. 2.8(b)) narrows the design space to a smaller range of acceptable L_{SDE} values for keeping I_{min} below the LSTP I_{OFF} target (10 pA/μm) while also keeping I_{ON} as high as possible. Another way of looking at this is to state that N_{SDE} must be as high as possible to keep I_{ON} as high as possible, but this comes at the cost of reducing L_{SDE} as L_G is scaled in order to keep I_{min} below the leakage specification, due to SCE and BTBT. Eventually a point will be reached where the SDE junction

cannot be made abrupt enough in practice to support both a high N_{SDE} and a small L_{SDE}, and the DSS MOSFET will no longer be suitable for LSTP applications. This reduction in L_{SDE} design space is much larger for NiSi than $ErSi_{1.7}$, since the lower electron SBH and higher hole SBH for $ErSi_{1.7}$ simultaneously improves I_{ON} and I_{min}, respectively. The I_{ON} improvement is due simply to the increase in source-body transmission through and over the smaller electron SB. The I_{min} improvement, however, is due to both the larger hole SBH and the lower electron SBH. The effect of the larger hole SBH is as described previously for Fig. 2.3. The effect of the smaller electron SBH is the ability to maintain a PD SDE region for smaller N_{SDE} values, which both reduces I_{min} (relative to FD SDE or PD SDE but with a smaller quasineutral region, as would be the case for NiSi) and SCE, thus enabling a larger design space for L_{SDE} as L_G is scaled, which Fig. 2.8(a) shows. Therefore, at ultimately small scales, near-midgap silicides are not competitive with near-band-edge silicides.

A general reduction in I_{ON} with L_G scaling is also seen in Fig. 2.8(a), which is due specifically to the reduction in optimal N_{SDE} from the onset of SCE and BTBT and therefore an increase in contact resistance R_c. Reducing *EOT* will improve I_{ON}, but there is still a net reduction in I_{ON} with L_G scaling and zero improvement in the design space, as Fig. 2.8(a) shows. Thus the effect of *EOT* scaling is only an increase in the inversion charge and not an improvement in gate modulation of the source-side SB, which is already made very small by the SDE region.

2.2.5 Effect of V_{DD} on I_{ON}

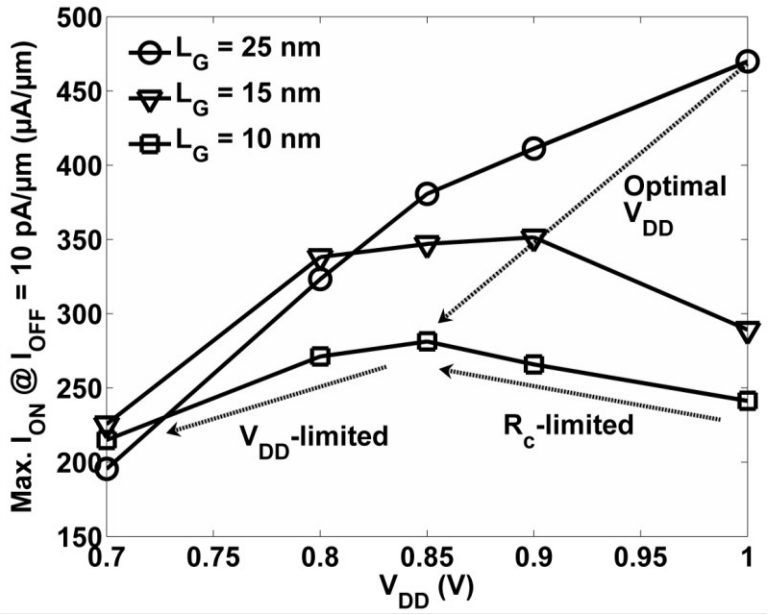

Fig. 2.9. Maximum LSTP I_{ON} vs. V_{DD} for double-gate FETs with NiSi MSD regions, for various L_G. In all cases, EOT = 1 nm, t_{body} = 2*L_G/3, L_{SDE} = 5 nm, and L_{un} 10 nm.

Another approach to reduce I_{min} and simultaneously increase I_{ON} is to reduce V_{DD}. This introduces a new trade-off, whereby reducing V_{DD} (which reduces I_{min}) permits an increase in N_{SDE} and therefore a reduction in R_c. The lower R_c afforded by the reduced V_{DD} would actually increase I_{ON} for the case of R_c-limited on-state conductance, although if V_{DD} is too low, then a net

reduction in I_{ON} results as the on-state conductance becomes voltage-limited rather than contact-limited. This is demonstrated in Fig. 2.9 for NiSi. It is important, and very interesting, to point out from these results that V_{DD} scaling in contact-limited structures is not arbitrary, but instead necessary in order to maximize I_{ON} for LSTP design.

2.2.6 LSTP Performance Comparison of DSS and RSD Structures

A fundamental limitation of DSS MOSFETs for LSTP design is evident. The SDE region must serve dual roles, which corresponds to a tradeoff between I_{OFF} and I_{ON} in the form of I_{min} vs. R_c. This tradeoff is exacerbated as L_G is scaled due to a reduction in the contact area (*i.e.*, t_{body} scales with L_G) and in the optimal N_{SDE}, so that on-state conductance becomes increasingly contact-limited at progressively smaller scales. Even when the SDE design parameters and V_{DD} are optimized, there remains a clear advantage of using low-barrier Schottky junctions (achieved with materials such as $ErSi_{1.7}$ for n-channel FETs or PtSi for p-channel FETs, or by using the aforementioned interface passivation techniques [17-21] in addition to dopants) for improved design space, and therefore immunity to process-induced variations, for LSTP devices. This is due specifically to the contact-limited nature of such devices. As a result, the potential for reduced source/drain series resistance in MSD MOSFETs is undermined by increased R_c, even under optimal design conditions.

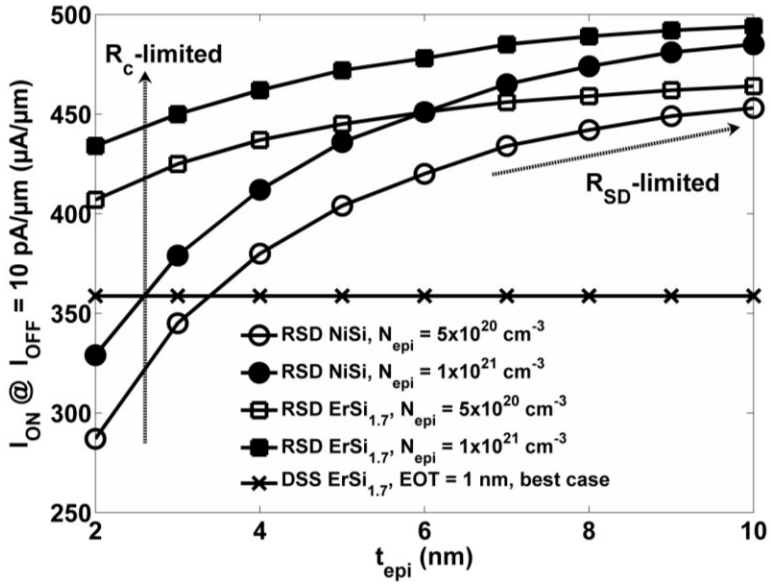

Fig. 2.10. I_{ON} vs. t_{epi} for RSD NiSi and $ErSi_{1.7}$ contacts and varying N_{epi}. L_G = 10 nm, L_{sp} = 10 nm, L_{SDE} = 5 nm, t_{body} = 6.67 nm, EOT = 1 nm, N_{SDE} = 1x10^{19} cm^{-3}, and V_{DD} = 1 V. The best case condition shown for DSS with $ErSi_{1.7}$ MSD is L_{un} = 10 nm, N_{SDE} = 1x10^{20} cm^{-3}, and V_{DD} = 0.9 V.

The simplest approach to improve R_c and therefore I_{ON} within a given material system is to increase the contact area. For a fixed value of t_{body}, this leads to the RSD structure (Fig. 2.1(b)). There are advantages to this structure that extend beyond increased contact area, though. Since the source/drain region is extended into two regions – the SDE region and the RSD region – low doping and high doping can be used in these regions, respectively, to simultaneously reduce I_{min} and R_c. Since the SDE region is no longer partially depleted by the SB, BTBT leakage will be

higher for the same N_{SDE} in the RSD structure (Fig. 2.6), and so a lower N_{SDE} is required to keep $I_{min} \leq I_{OFF}$. Here N_{SDE} is set conservatively to $1x10^{19}$ cm^{-3}, given the BTBT results in Fig. 2.5 and noting that SB/SDE tunneling leakage is non-existent in the RSD structure.

Although spreading resistance R_{sp} becomes a significant factor in the RSD structure, increasing t_{epi} reduces its effect due to a reduction in current crowding. This is shown in Fig. 2.10, which also shows that for any t_{epi}, I_{ON} in the RSD structure is significantly larger for the same SBH due to the increased contact area and therefore reduced R_c. For large enough t_{epi}, reducing the SBH has little effect on increasing I_{ON}, and in this regime, a larger increase in I_{ON} is achieved by increasing N_{epi}. This suggests that series resistance within the RSD and SDE regions is limiting the gains achievable by reducing R_{sp} with increasing t_{epi}, and so as expected, the optimized RSD structure is R_{SD} limited. Nevertheless, the RSD structure obviates the need for low barrier contacts due to its higher contact area and capacity to use high doping in the RSD regions while keeping $I_{min} < I_{OFF}$. It is specifically this advantage that results in the RSD structure being superior for LSTP design.

2.3 HP Design Study

This study begins with a 2D double-gate (DG) DSS NMOS structure, as shown in Fig. 2.11. L_G = 10 nm, $t_{body} = L_{sp} = L_{SDE}$ = 7 nm, V_{DD} = 1 V, I_{OFF} = 100 nA/μm, and t_{ox} = 1 nm. L_{sp} is the gate underlap to the source/drain SB junctions (also the sidewall spacer thickness if one neglects lateral silicidation, as is done here), where the sidewall spacer is made of silicon nitride [55], t_{flare} is the amount by which the source/drain silicide regions adjacent to the sidewall spacer flare out from the fin structure, and all other terms have their usual meaning. The metal gate height t_{gate} = 20 nm, the body doping is $1x10^{15}$ cm^{-3} p-type, and $N_{SDE} = 3x10^{20}$ cm^{-3}, while L_{SDE} is varied. The SB height (SBH) at the M-S interface is set to 0.1 eV in all cases simulated here (including the 3D structures shown later), which is reasonable considering reported data on dopant segregation and interface passivation by Group-VI species [17], [56], [57] as well as the results in Chapters 4 and 6. The silicide workfunction ϕ_M is varied independently from that of the SB contacts.

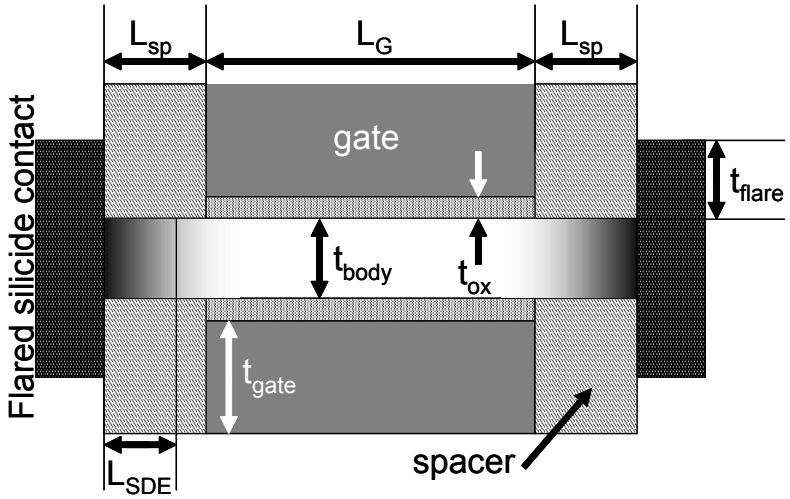

Fig. 2.11. Schematic cross-section of the 2D double gate DSS structure modeled here.

NiSi is the primary silicide material simulated here, given its ubiquity in modern CMOS processing. Reported data on NiSi ϕ_M varies from ~4.6 eV to 4.7 eV [58]-[60], although it is not

clear in those studies what value is assumed for the electron affinity of silicon as a reference for extracting ϕ_M. (Sentaurus Device assumes the electron affinity of silicon to be 4.07 eV.) For the sake of simplicity, ϕ_M is assumed here to be 4.72 eV, which correlates well with the reported electron SBH of 0.65 eV for NiSi [17].

2.3.1 Effect of Silicide Gating on DSS FinFET Performance

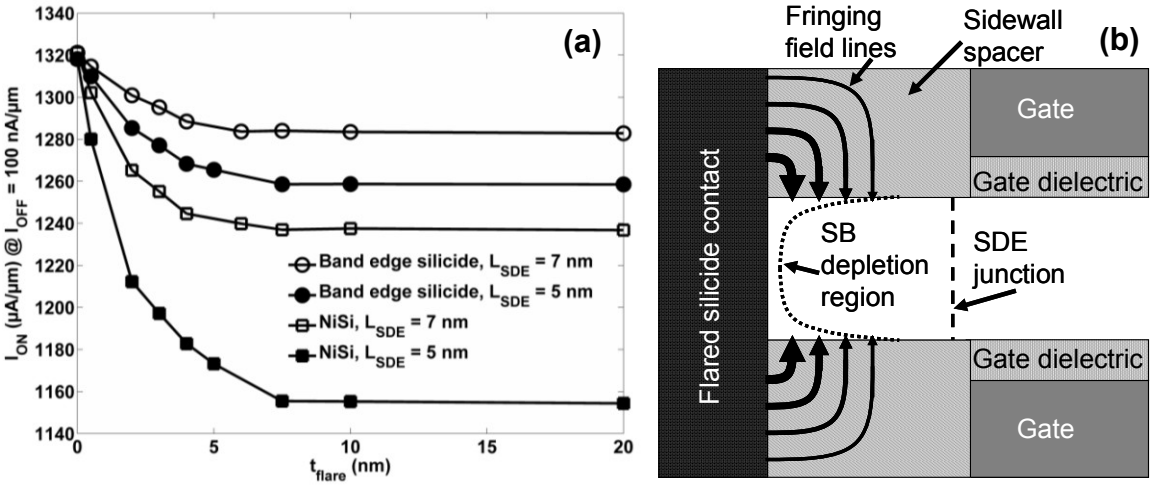

Fig. 2.12. (a) effect of silicide gating on DSS performance for the device structure shown in Fig. 2.11, with two different silicides and (b) illustration of the silicide gating effect. Fringing field lines from the silicide traverse the sidewall spacer and terminate on the SDE region, depleting carriers away from the SB contacts and increasing the contact resistance as well as the series resistance in the SDE region.

Fig. 2.12(a) shows the effect of t_{flare} on I_{ON} for NiSi and also for a conduction band edge silicide (with ϕ_M = 4.07 eV). For t_{flare} = 0 nm, both silicides exhibit the same I_{ON}, which degrades as t_{flare} increases. This is due to the fringing field effect of the silicide depleting carriers from the heavily-doped SDE region, which has a lower workfunction than the silicide (i.e., the silicide gating effect), as Fig. 2.12(b) illustrates. For NiSi, 1 nm of Ni reacts with 1.84 nm of Si to result in 2.22 nm of NiSi [31], so that silicidation of a straight fin results in some NiSi flaring out from the fin due to volume expansion. For 7 nm fin width, this flaring is ~0.72 nm on either side of the fin which results in I_{ON} ~ 50 μA/μm lower than the ideal value for L_{SDE} = 5 nm, according to Fig. 2.12(a). This effect can be minimized by using a band-edge or near-band-edge silicide, as Fig. 2.12(a) shows. Another approach is to reduce the SDE abruptness (*i.e.,* employ larger L_{SDE}), as can be seen from Fig. 2.12(a) by comparing the curves for L_{SDE} = 5 *vs.* 7 nm. In this case, the average doping within the SDE region under the sidewall spacer is higher, resulting in less overall SDE depletion and therefore lower series resistance. As Fig. 2.12(a) shows, the silicide gating effect saturates at t_{flare} ~ 7.5 nm, mostly independent of ϕ_M or L_{SDE}, for the design parameters used here. By reducing L_{sp}, the additional fringing fields from the silicide generated by increasing t_{flare} are screened out by the gate electrode and so never terminate on the SDE regions. As a result, the silicide gating effect will saturate at a lower t_{flare} and the maximum I_{ON} reduction is reduced, as Fig. 2.13 shows. This permits larger t_{flare} (i.e., thicker silicides) with less performance penalty relative to t_{flare} = 0 for a given L_{sp}. Although there is a natural tradeoff here between the impact of L_{sp} on short channel effects (SCE), I_{ON} and C_{ov}, there is a more important

correlation to metal deposition technology. As mentioned earlier, a potential challenge to aggressively scaled DSS FinFET fabrication is the need to deposit thin metals. If this cannot be met, then one of three design possibilities exists: L_{sp} must be increased to compensate for lateral silicidation, or an epitaxial layer must be added to the fin source/drain regions to allow for a thicker silicide with reduced or zero lateral silicidation, or both.

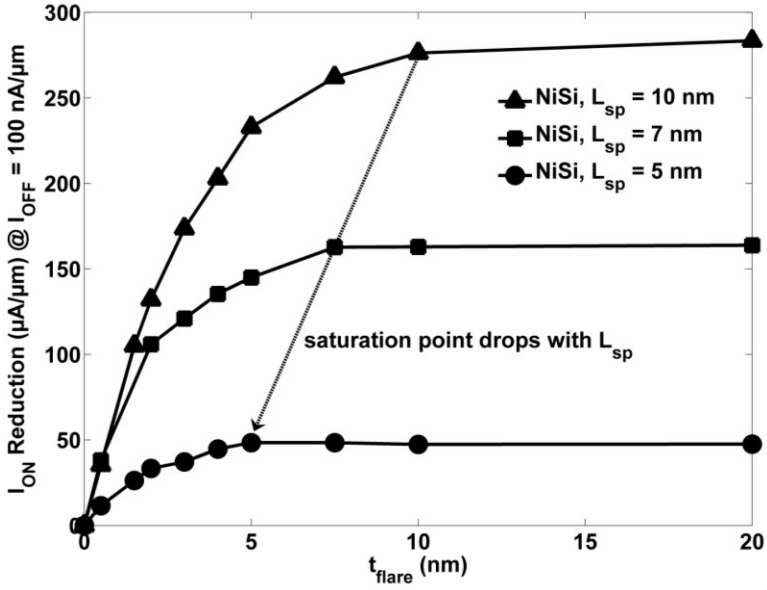

Fig. 2.13. Effect of L_{sp} on silicide gating, for the device structure shown in Fig. 2.11 with ϕ_M = 4.72 eV (NiSi) and L_{SDE} = 5 nm.

If L_{sp} is increased to offset lateral silicidation, then forming abrupt, heavily-doped SDE regions becomes more difficult, regardless of whether dopant pile-up or implant-to-silicide (ITS) is utilized. For dopant pile-up, N_{SDE} drops as the silicidation front progresses [50]. Thus, increasing L_{sp} will mean the silicidation process must progress further laterally for the same L_{SDE} and contact-to-gate edge spacing, thereby reducing N_{SDE} and increasing contact resistance R_c. However, this lateral silicidation encases some silicide between the top and bottom sidewall spacers, resulting in zero localized t_{flare} and therefore zero silicide gating effect from the silicide region closest to the SDE region. If ITS is utilized instead of dopant pile-up, the portion of the silicide overlapped by the sidewall spacer ends up not being exposed to the SDE implant, resulting in a difference in silicide grain size in the implanted region (smaller grains due to implant damage) and the spacer-protected region (larger grains). It has been shown [61] that this grain size affects the diffusion and amount of barrier lowering achieved with low work function metals in NiSi, so it stands to reason that the effect on dopant diffusion within silicides is similar. This may influence the optimal ITS implant and anneal conditions. The optimal lateral silicidation for either approach will therefore be determined by the tradeoff between R_c (due to dopant segregation) and silicide gating, which itself affects R_c.

If an epitaxial layer is added to the fin source/drain regions to minimize lateral silicidation, then t_{flare} increases, reducing I_{ON} through silicide gating. This can be mitigated by reducing L_{sp} as mentioned, but at the cost of higher C_{ov}. As a result, metal deposition technology will define the boundary conditions for L_{sp}, t_{flare}, *etc.* in a DSS process. In further analyses herein on DSS FinFETs, it is assumed that thin metal films can be deposited reliably and uniformly to fully silicide the fin source/drain regions without negatively influencing the process of forming

heavily-doped and moderately abrupt SDE regions. This leads to the 3D DSS structure in Fig. 2.14. The geometry, doping, *etc.* are the same as for the 2D DSS structure in Fig. 1, but with the silicide source/drain regions extending outward from the sidewall spacers by 30 nm. Volume expansion after silicidation is considered, resulting in a small t_{flare} of 0.72 nm (i.e., NiSi is assumed). Also, the fin source/drain regions are strapped with Metal 1 (M1) layer bars, with a width of 10 nm and a pitch of 60 nm. (Since t_{flare} must be kept to a minimum in an optimized DSS FinFET, the source/drain regions must be kept narrow and so they are connected together using M1 contact bars.) The volume not filled by the FinFET structure or the M1 bars is filled by an inter-layer dielectric (ILD).

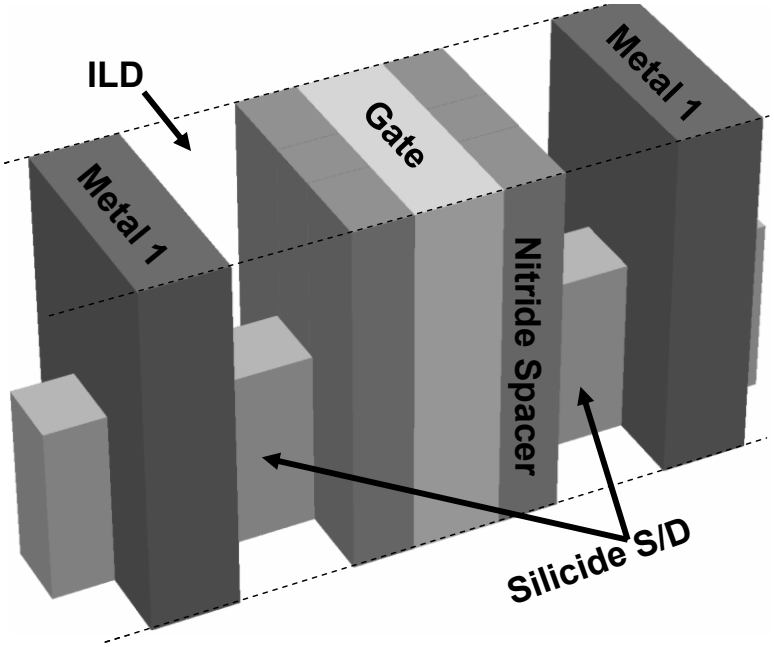

Fig. 2.14. 3D illustration of the DSS FinFET modeled in this study.

2.3.2 3-D Contact Optimization for RSD FinFETs

Source/drain and contact optimization involves different tradeoffs for RSD FinFETs. A well-known challenge is the formation of uniformly doped source/drain regions throughout the height of the fin to achieve low source/drain series resistance R_{SD}, given implant shadowing and dose loss effects [62], [63]. Plasma doping [64], [65] has been proposed as an alternative technique to achieve uniform fin doping, although plasma-induced damage poses challenges for balancing recrystallization and dopant diffusion to minimize R_{SD}. Gas-phase doping is another option, which relies on the adsorption of dopant carrier gases onto the silicon surface in a vacuum environment [66], [67]. This is a simple and damage-free process, although ultraclean surfaces (*i.e.*, no native oxide) are required and dose controllability is limited by the ability to quickly and uniformly fill and evacuate the process chamber of dopant gases. In the analysis to follow, the RSD structure is assumed to be ideal in that it has uniformly doped and damage-free source/drain regions.

The conventional approach for fabricating RSD FinFETs is to form an epitaxial RSD region around the fin source/drain region and then partially silicide this epitaxial region, thus forming a wrapped contact (WC) structure [68], [69]. This is very similar to a DSS FinFET with an epitaxy-induced t_{flare}, although here the source/drain region is not fully silicided. In this case, the optimal I_{ON} is affected by how much the silicide consumes the RSD region, which affects current crowding within the source/drain regions [69], [70]. A non-zero RSD thickness (t_{RSD}) increases C_{ov}, which is also increased by the metal bar that would be needed to strap the fins in a multi-fin device. It has been argued that [69], [71], especially for small fin pitches (FP), fins strapped by lateral epitaxial growth in the source/drain regions offer lower C_{ov} due to the fact that the vias used to access the source/drain regions have a smaller sidewall surface area than the metal bars that would otherwise be needed to strap the fins. As a result, RSD FinFETs with minimal C_{ov} require a top-contacted (TC) approach, since the lateral epitaxial fin strapping eliminates silicide access to the source/drain sidewalls. Provided the source/drain doping is high enough (~$1x10^{20}$ cm^{-3}) and, more importantly, the silicide contact SBH is low enough (~0.1 eV), the difference in I_{ON} between TC and WC RSD FinFETs is negligible (<3% according to simulations, not shown here).

The question then arises: what is the best way to form a TC structure? Considering that silicidation consumes silicon, a "true" TC RSD structure would require the epitaxial growth of silicon vertically on top of the source/drain regions, in order for the silicide-silicon contact interface to be co-planar with the top of the fin. This is the "recessed silicide" structure shown in Fig. 2.15(a), for which C_{ov} increases with silicide thickness. Alternatively, if no epitaxial silicon is grown prior to silicidation, the "embedded silicide" structure shown in Fig. 2.15(b) would result. Note that only a fraction of the silicide extends up above the original source/drain surface (due to silicon consumption and volume expansion), so that it has lower C_{ov}. The top portion of the transistor is essentially a DSS structure with large t_{flare}, while the bottom portion is a TC RSD structure. The problem here is three-fold. First is silicide gating in the top portion of the source/drain fin regions, as discussed previously for the DSS structure. Second is current crowding in the bottom portion of the source/drain fin regions as the fin height H_{fin} is reduced and/or as the silicide thickness is increased (while at the same time a larger fraction of the source/drain region becomes a DSS structure with large t_{flare}, so I_{ON} will drop sharply for small H_{fin}). Third is that the silicide is formed downward from the top surface rather than from the sides of the fin: the silicide junction is rounded, so the distance between the contact and the gate edge increases moving downward from the top of the fin. (The silicide junction is assumed to be perfectly square in this study, so this effect is ignored.) The recessed silicide structure does not have these problems, but suffers from higher C_{ov} due to the larger combined RSD and silicide area abutting the sidewall spacer. Also, since only the top of the source/drain regions are contacted, I_{ON} saturates due to the increased voltage drop toward the bottom of the source/drain regions as H_{fin} is increased, unless the source/drain regions are very heavily and uniformly doped (as is assumed here).

The doping profiles and device geometry for the RSD FinFETs in Fig. 2.15 are the same as for the basic 2D DSS structure in Fig. 2.11, except that the source/drain regions extending outward from the sidewall spacers (which includes the fin and flared source/drain regions) are comprised of uniformly doped silicon (N_{SDE} = $3x10^{20}$ cm^{-3}) rather than silicide, with a length of 30 nm. For both the recessed and embedded silicide structures, the inner silicide edge directly abuts the sidewall spacer and, for the embedded silicide structure, it is assumed that no lateral silicidation under the spacer takes place.

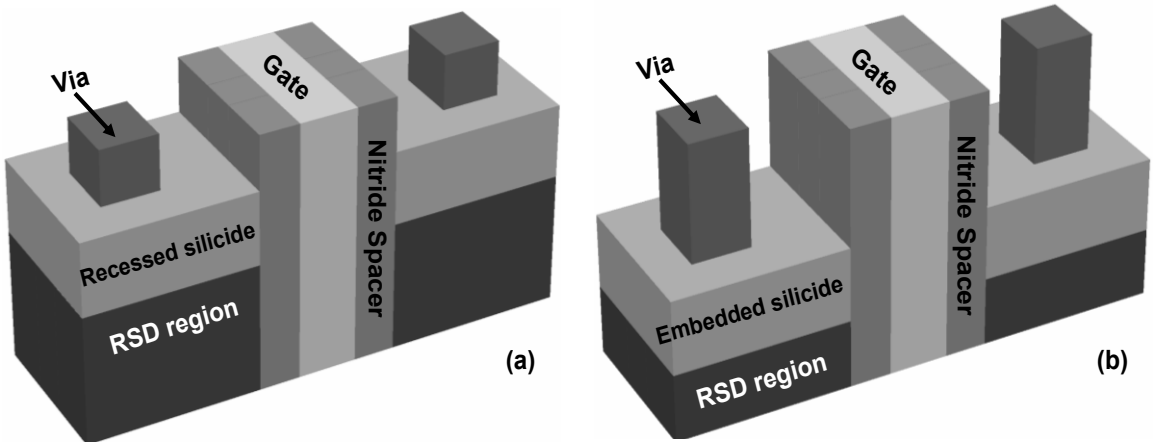

Fig. 2.15. 3D illustrations of the (a) recessed silicide and (b) embedded silicide RSD FinFETs modeled in this study. In (a), the recessed silicide is placed on top of the source/drain region, resulting in the height of the silicon RSD region equaling that of the fin under the gate electrode. In (b), the embedded silicide consumes a portion of the source/drain region, resulting in the height of the silicon RSD region being less than that of the fin under the gate electrode.

For the comparative study of DSS *vs.* RSD FinFETs in the next section, the gate electrode wraps around the fin, forming a tri-gate (TG) structure as opposed to a double gate (DG) structure. This is to reflect realistic 3D FinFET fabrication at aggressive dimensions, for which topography should be limited. The 2007 ITRS Roadmap [14] is used as a guideline for defining via pitch, silicide thickness, gate height, and ILD permittivity: The vias are 10 nm x 10 nm, with a pitch of 60 nm, with the top of the via coplanar with the top of the gate electrode and gate-sidewall spacers. The silicide thickness is 12 nm and the gate height (extending upwards from the top of the fin) is 20 nm. The gate "height" extending outward from the fin sidewalls is limited by the fin spacing, and is defined here as half of the fin spacing since a single FinFET is being modeled. The ILD region, which fills the open space around the vias, has a relative permittivity $\varepsilon_r = 2.3$. This value is taken as the average of the range of values speculated in the ITRS Roadmap for 10 nm L_G design. FP is varied from 14 nm (7 nm fin and 7 nm fin spacing) to 57 nm (7 nm fin and 50 nm fin spacing), while H_{fin} is varied from 10 nm to 21 nm (*i.e.*, the maximum fin aspect ratio is 3:1).

2.3.3 DSS vs. RSD FinFET AC Performance

The dependence of intrinsic delay on FP is shown in Fig. 2.16(a) for the DSS and RSD structures, for ϕ_M values of 4.07 (band edge silicide) and 4.72 eV (NiSi). Here the intrinsic delay is defined as $C_{GG}*V_{DD}/I_{eff}$, where C_{GG} is the modeled input capacitance for the gate electrode and I_{eff} is the average of I_{DS} for $V_{DS} = 0.5*V_{DD}$ with $V_{GS} = V_{DD}$ and I_{DS} for $V_{DS} = V_{DD}$ with $V_{GS} = 0.5*V_{DD}$. What is most evident from Fig. 2.16(a) is that the dependence of delay on FP is much lower for the DSS structure. This is because the average distance from the gate electrode to a fringing capacitance element (*e.g.*, M1, via, or RSD regions) is larger for the DSS structure. Since capacitance is inversely proportional to distance and directly proportional to area (which is

determined by FP and device height), the slope of the delay curve for the DSS structure will be lower. This indicates a much larger AC design space for DSS FinFETs over RSD FinFETs with respect to FP, which is very interesting because in [70] it was shown that RSD structures have a larger DC design space. However, Fig. 2.16(a) also shows that the RSD structure with embedded silicide achieves the lowest delay at small FP. This is because the RSD regions contribute less to C_{GG} than the vias or M1 bars, for small FP, as mentioned before and in [69], [71]. Since the M1 bars have more lateral surface area in this regime, the C_{GG} advantage for the DSS structure diminishes (Fig. 2.16(b)). The fact that delay is lower and C_{GG} is about the same for DSS and RSD with embedded silicide for small FP suggests that the RSD structure has some I_{eff} advantage, despite the silicide gating effect (which is stronger for the embedded silicide structure) in the top portion of the source/drain fin regions. Although not shown here explicitly, the RSD region has a gating effect of its own, similar to silicide gating but with an opposite effect: if the flared RSD regions are very heavily doped, they have low work function resulting in some electron accumulation in the decaying SDE profile regions under the sidewall spacer (throughout the entire height of the SDE for the recessed silicide structure, but only in the bottom portion of the SDE for embedded silicide).

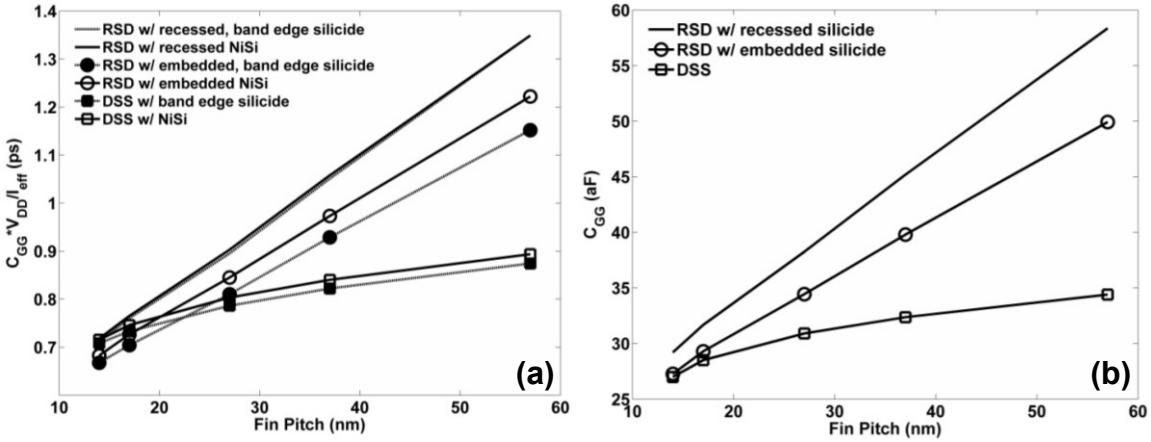

Fig. 2.16. (a) $C_{GG}*V_{DD}/I_{eff}$ vs. FP and (b) corresponding C_{GG} vs. FP for the 3D DSS and RSD FinFETs for different silicide workfunctions and H_{fin} = 21 nm.

If one considers the effect of H_{fin} on delay, then the situation becomes more complicated. Recall that, as H_{fin} is reduced, the RSD FinFET with embedded silicide experiences degradation in I_{ON} due to a larger fraction of the source/drain region being a DSS structure with large t_{flare}, as well as increased current crowding in the bottom doped source/drain region. Fig. 2.17 shows the strong effect that this has on delay, to the point where if H_{fin} is small enough (< 13 nm in Fig. 2.17) then the optimal source/drain design is neither the RSD structure with embedded silicide nor the DSS structure (due to silicide gating and gate-to-M1 fringing capacitance), but instead the RSD structure with recessed silicide. However, this is the case only for NiSi. If ϕ_M is small enough, then for H_{fin} < 13 nm the DSS structure achieves the lowest delay.

What this means is that the optimal source/drain design depends on FP, H_{fin}, and ϕ_M. For ϕ_M ~ midgap, DSS is faster for large H_{fin} and FP, while RSD with embedded silicide is faster for large H_{fin} and small FP, and finally RSD with recessed silicide is faster for small H_{fin} and FP. As ϕ_M drops to band-edge or near-band-edge values, the DSS structure is fastest in every case except for large H_{fin} and small FP. In practice, the vertical doping profile may be non-uniform, so the delay may increase with H_{fin} for very large H_{fin}, as shown in [62] for H_{fin} > 70 nm. If N_{SDE} drops

at the bottom of the fin as H_{fin} is increased, I_{eff} will saturate or at least taper off as H_{fin} increases. Also, since the gate area and therefore C_{GG} increases with H_{fin}, the combined increase in C_{GG} and saturation in I_{eff} will increase delay. If anything, this points to a potential integration advantage for DSS structures, because dopant diffusion in silicides is very fast [56] and so a vertically uniform SDE profile may be more easily achievable with an ITS process.

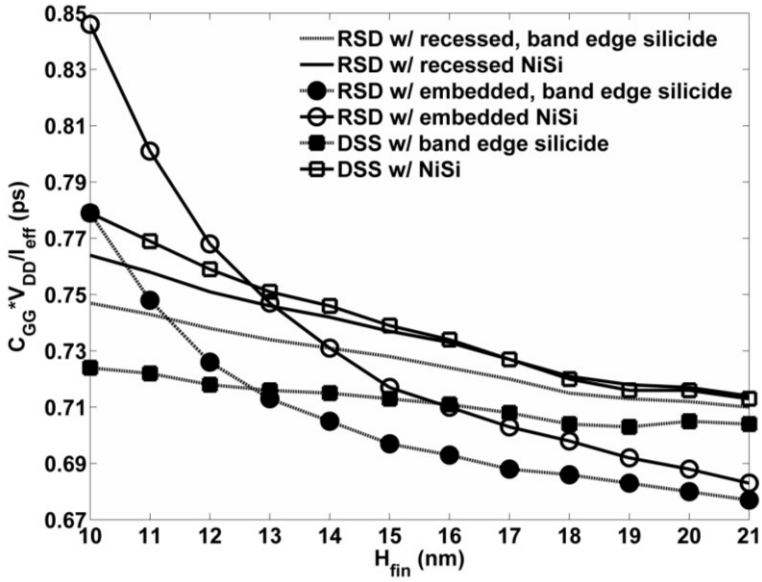

Fig. 2.17. $C_{GG}*V_{DD}/I_{eff}$ vs. H_{fin} with FP = 14 nm for the 3D DSS and RSD FinFETs and ϕ_M = 4.07 eV and 4.72 eV (NiSi).

The C_{GG} vs. H_{fin} dependence (not shown) is linear for the devices simulated in Fig. 2.17 (*i.e.,* FP = 14 nm), with the only difference being the y-intercept. This y-intercept is C_{GG0}, or the contribution of the M1/via region and the silicide above the top of the fin to C_{GG}. The DSS structure exhibits the lowest C_{GG0}, at 5.86 aF. Next is the RSD structure with embedded silicide, at 6.36 aF. Although the via cross-sectional area in the RSD structure is lower compared to the M1 strapping bar in the DSS structure, this is more than offset by the small amount of silicide extending up from the top of the fin due to volume expansion during silicidation. This is ultimately a moot point, though, since for small H_{fin} the embedded silicide structure experiences a severe delay penalty due to silicide gating reducing I_{eff}. For the recessed silicide structure, C_{GG0} is much higher at 8.07 aF and is why delay for this structure suffers for low H_{fin} and ϕ_M.

Surely, one can consider scaling down the silicide thickness even more aggressively than the ITRS Roadmap suggests (12 nm). This would not affect the DSS structure (which already has a fully silicided source/drain), but it would affect C_{GG} for the RSD structure with recessed silicide and I_{eff} for the RSD structure with embedded silicide. Fig. 2.18 shows the impact of scaling the silicide to 8 nm and 4 nm. Both the delay and sensitivity to H_{fin} drop with silicide thickness for the RSD structures, most notably for the embedded silicide structure for which silicide gating has the most influence. Also, the performance of both structures over the H_{fin} range converges as the silicide thickness drops, since they each approach the same idealized TC RSD structure.

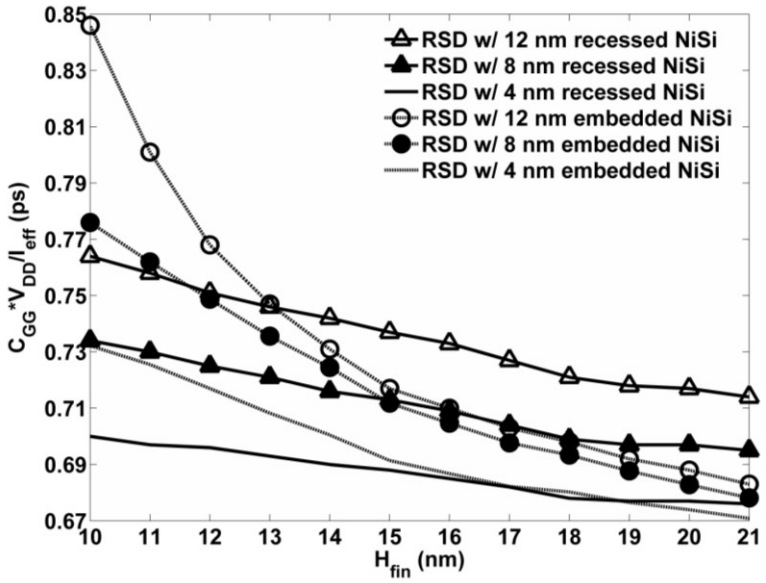

Fig. 2.18. C_{GG}*V_{DD}/I_{eff} vs. H_{fin} with FP = 14 nm for the 3D RSD FinFETs, with NiSi thicknesses of 4 nm, 8 nm, and 12 nm.

It is worth noting that scaling down the silicide thickness is a very difficult task. For example, 12 nm of NiSi would require an as-deposited Ni thickness of 5.4 nm. For 8 nm and then 4 nm of NiSi, the as-deposited Ni thickness drops to 3.6 nm and 1.8 nm, respectively, which is only a few atomic layers. It has been reported for $ErSi_{1.7}$ [38] that the SBH (and therefore contact resistance) increases significantly if the silicide thickness drops below 12 nm, since below this thickness the number of metal-induced gap states (MIGS) needed to pin the Fermi level to a low SBH condition has not yet saturated to its bulk value. (This effect is ignored in Fig. 2.18.) As a result, ultra-thin silicides will likely be impractical, and so the optimal FinFET structure which provides an AC performance advantage over all ranges of FP and H_{fin} with realistic silicide thicknesses will require a different source/drain geometry than that of either the conventional DSS or RSD structures.

2.3.4 Recessed Strap (RS) DSS FinFETs

Fig. 2.19 shows the recessed strap (RS) DSS FinFET, while Fig. 2.20 illustrates the main process steps used to achieve this structure. The ILD region used here is the same as for the RSD and conventional DSS structures. The use of lateral epitaxial fin strapping, just like the RSD structure, permits the use of vias rather than M1 bars for lower delay in the small-FP regime. At the same time, recessing the strapping region (by a recess depth D_r) results in lower C_{GG} and reduced silicide gating, leading to reduced delay dependence on FP and H_{fin}, just like the conventional DSS structure. Furthermore, M1 fin strapping for conventional DSS FinFETs results in fringing capacitance C_{fr} between the gate and M1 bar for each fin connected in parallel. With lateral epitaxial fin strapping, via pitches larger than FP can be utilized (Fig. 2.20(d)) to reduce C_{fr}. This is an advantage that both the RSD and the RS DSS FinFET has over the conventional DSS FinFET structure, although here a worst-case scenario is assumed in which each device experiences the full C_{fr} penalty from two vias.

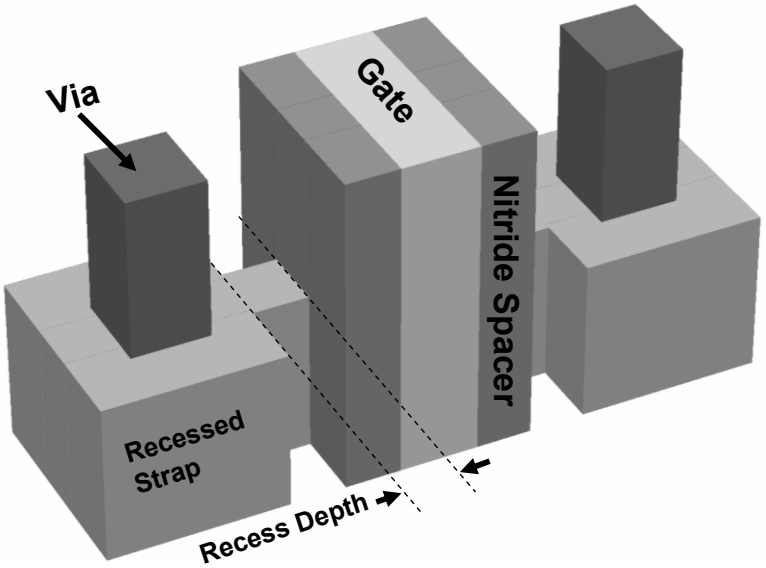

Fig. 2.19. 3D illustration of the RS DSS structure modeled in this study.

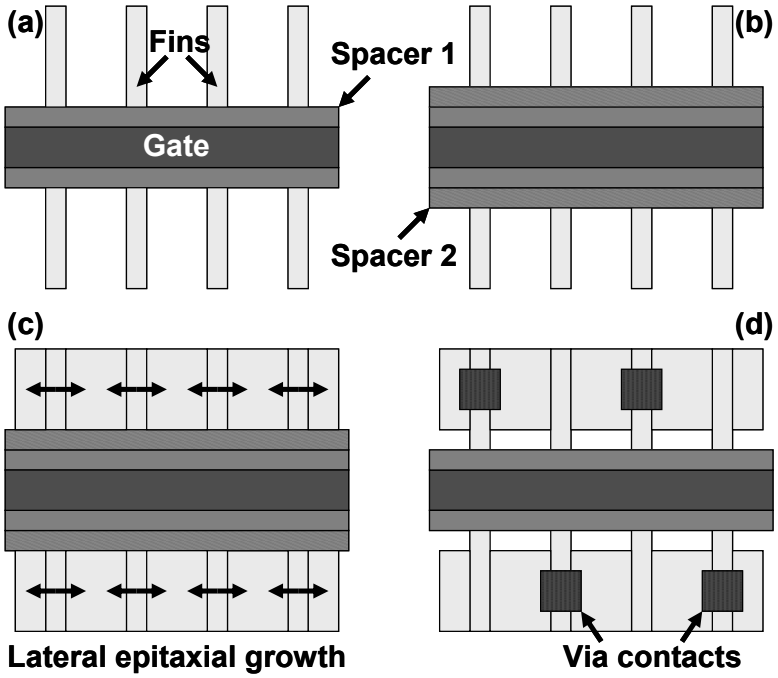

Fig. 2.20. Main process flow for RS DSS FinFETs (top-down view). After forming the fins, gate, and sidewall spacers (a), a second sidewall spacer is formed (b) which defines the recess depth of the fin strapping regions. Then (c) lateral epitaxial growth is performed to strap the fins. The second spacer is then etched away (d), followed by silicidation and via formation. Instead of two separate spacers, one can alternatively use a single, wider spacer and then etch it back partially after strapping the fins.

The end result is a FinFET source/drain architecture that combines the merits of both DSS and RSD FinFETs to result in equivalent or improved AC performance over all ranges of FP and H_{fin}, as Fig. 2.21 shows. The use of ILD to separate the recessed strap region from the nitride spacer results in the RS DSS structure having the lowest C_{GG0} among the structures modeled here, at 5.3 aF.

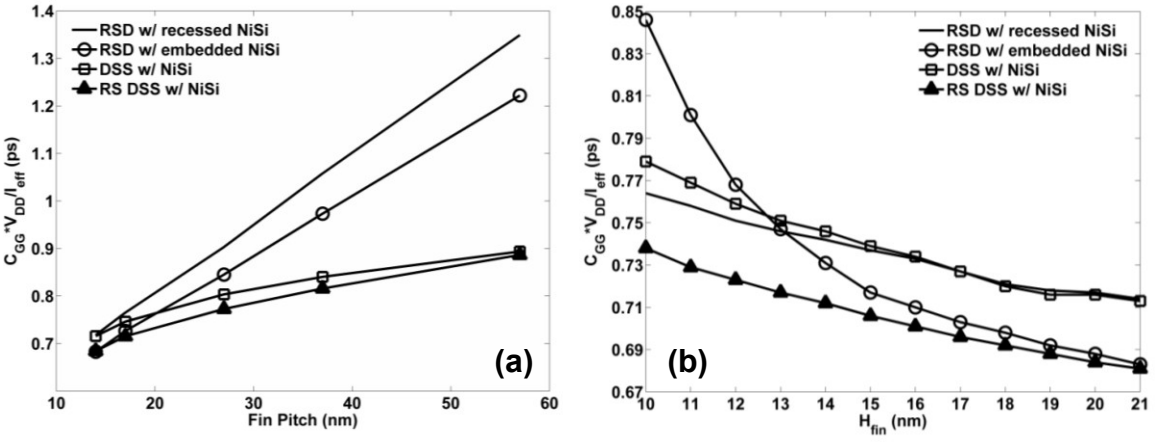

Fig. 2.21. (a) C_{GG}*V_{DD}/I_{eff} vs. FP with H_{fin} = 21 nm and (b) C_{GG}*V_{DD}/I_{eff} vs. H_{fin} with FP = 14 nm for the RSD, DSS, and RS DSS FinFETs with NiSi contacts. For the RS DSS FinFET, D_r = 7 nm.

One may argue that a simpler approach than the RS DSS FinFET would be to use the RSD FinFET with embedded or recessed silicide, but with a larger L_{sp} to reduce C_{GG}. Fig. 2.22(a) shows this performance comparison, where L_{sp} = 7 (default) and 14 nm for the RSD FinFETs (this is equal to L_{sp} = 7 nm plus D_r = 7 nm for the RS DSS FinFET). Also, L_{SDE} = 14 nm for the RSD structure (with L_{sp} = 14 nm) to keep the effective channel length L_{eff} the same (L_{eff} is defined here as $L_G + 2*(L_{sp} - L_{SDE})$. As Fig. 2.22(a) shows, this approach (increasing L_{sp} for the RSD FinFET) does not improve upon the RS DSS approach. Although the RSD delay dependence (with L_{sp} = 14 nm) on FP is improved compared to L_{sp} = 7 nm and the delay is smaller for large FP, the delay is actually higher for small FP. This is similar to but actually worse than for the conventional DSS FinFET, as Fig. 2.22(a) also shows.

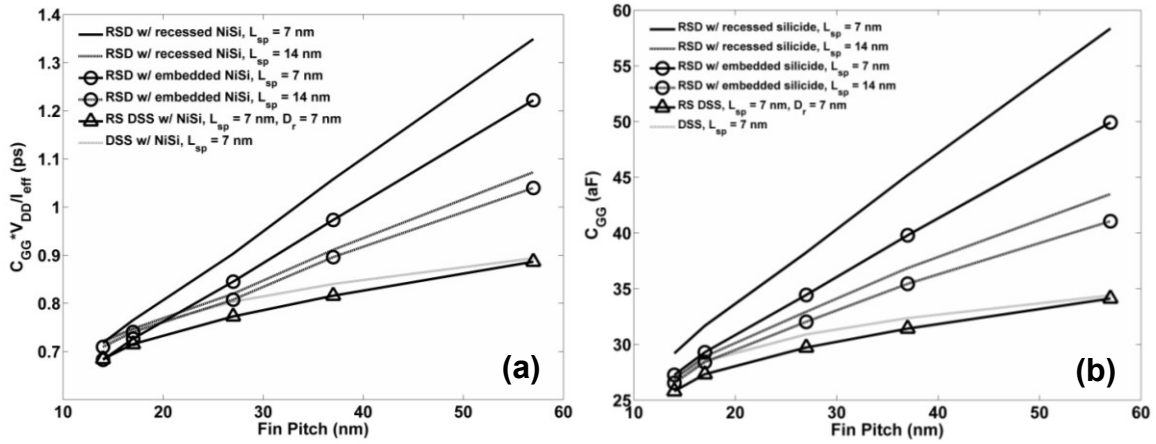

Fig. 2.22. (a) C_{GG}*V_{DD}/I_{eff} vs. FP with H_{fin} = 21 nm and (b) corresponding C_{GG} vs. FP for the 3D RSD, DSS, and RS DSS FinFETs with NiSi contacts, with L_{sp} = 7 and 14 nm for the RSD FinFETs.

The problem with the increased L_{sp} approach for the RSD structure is two-fold. First, since the extra 7 nm of spacer thickness in the RSD structure is provided by the spacer and not the ILD region, C_{GG} is larger compared to the RS DSS structure (Fig. 2.22(b)). Second, since the underlapped SDE region, which is defined by doped silicon, is longer in the RSD structure, source/drain series resistance will be higher compared to the RS DSS structure (neglecting the

effect of RSD gating). This is due in part to the longer source-to-drain contact spacing but also due to the gate fringing field having less effect over accumulating electrons over the entire length of the SDE region in the on-state.

Although not shown here, the I_{eff} penalty in moving from L_{sp} = 7 nm to L_{sp} = 14 nm ranges from ~2.5-6.3% for RSD with embedded silicide and ~5.8-7.4 % for RSD with recessed silicide. This is FP-dependent, being larger for small FP, due to a reduction in RSD volume with FP, which causes I_{eff} to drop due to the increase in R_{SD}. Especially for the embedded silicide RSD FinFET, the result is that for small FP, where C_{ov} to the RSD regions is already small (*i.e.*, C_{GG} is comparable for L_{sp} = 7 *vs.* 14 nm), the I_{eff} penalty increases delay. (This also explains the small kink in the delay curves in Fig. 2.22(a) for L_{sp} = 14 nm and FP < 27 nm, which is not observed for C_{GG} in Fig. 2.22(b)).

Another approach for large-L_{sp} RSD FinFETs is to use dual spacers, similar to Fig. 2.20(c), where the outer spacers have lower dielectric constant than the inner spacers. This would primarily serve to reduce C_{GG}, although I_{eff} would also drop. Two cases were simulated for the dual-spacer RSD FinFETs, with inner and outer spacers of equal length (7 nm): nitride inner spacers (ε_r = 7), and either oxide (ε_r = 3.9) or ILD (ε_r = 2.3) outer spacers, each for H_{fin} = 21 nm and FP = 14 and 57 nm, with embedded silicide contacts. For oxide outer spacers, $C_{GG}*V_{DD}/I_{eff}$ is 0.03 % lower and 2.03 % higher than for the RS DSS FinFET for FP = 14 and 57 nm, respectively; for ILD outer spacers, $C_{GG}*V_{DD}/I_{eff}$ is 1.68 % lower and 7.06 % lower than for the RS DSS FinFET for FP = 14 and 57 nm, respectively. Thus, if ε_r for the outer spacers is low enough, the dual-spacer RSD FinFET structure with embedded silicide achieves the lowest delay and has the lowest delay dependence on FP. (The H_{fin} dependence problem with embedded silicide RSD FinFETs would remain, though). However, there may be a process integration challenge here. For RS DSS FinFET fabrication, the outer spacer is etched away after strapping the fins but before forming the silicide. For the dual-spacer RSD FinFET, the second spacer must be in place as the silicide contact is formed; otherwise, the narrow SDE regions would be fully silicided, resulting in a RS DSS FinFET. This would not be a problem if the outer spacers were oxide – but then the dual-spacer RSD FinFET would have equivalent or worse performance as the RS DSS FinFET. For ILD outer spacers, the porous low-k material must be able to withstand the silicidation process without increasing ε_r [72]. This may be impractical for typical silicidation temperatures in the range 350 – 600 °C. Thus, adjusting L_{sp} in RSD FinFETs, either with single or dual spacers, is likely not the best approach for optimizing FinFET delay and, more fundamentally, source/drain design, under practical process conditions.

These problems are avoided in the RS DSS FinFET, albeit at the expense of a smaller DC design window [70]. It is also worth noting that the maximum D_r in the RS DSS FinFET will be constrained by the lithographic alignment tolerance which determines the minimum spacing between the inner via edge and the inner edge of the source/drain strapping region. Up to this point, D_r was set to 7 nm, so that the strap falls about halfway between the sidewall spacer edge and the inner via edge for the via size/pitch used here of 10/60 nm. Fig. 2.23 shows the dependence of delay on D_r for different values of FP. As already indicated in Figs. 2.21 and 2.22, a reduction in FP will reduce the delay. However, reducing FP also improves the D_r design space, since both C_{GG} and silicide gating scale with FP, permitting smaller D_r with reduced or zero performance penalty.

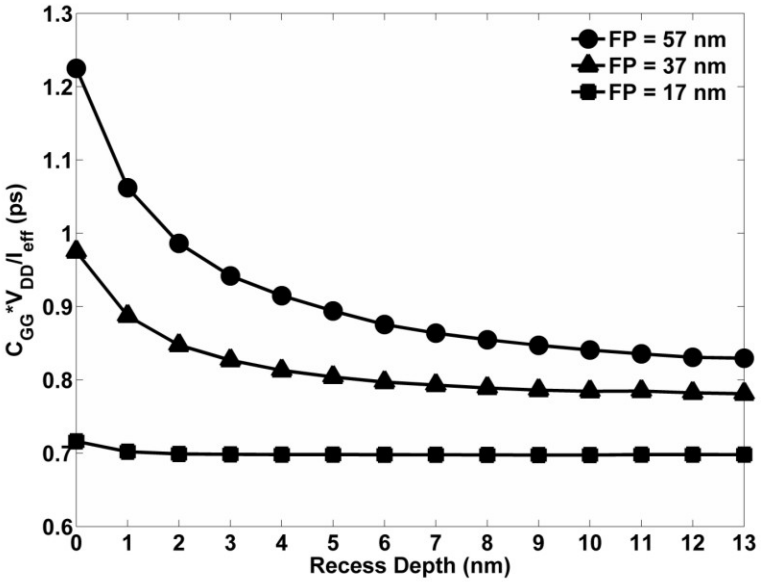

Fig. 2.23. C_{GG}*V_{DD}/I_{eff} vs. D_r with H_{fin} = 21 nm for the 3D RS DSS FinFET with NiSi contacts. D_r = 0 nm means the fin strapping layer abuts the sidewall spacer, while D_r = 13 nm means the inner edge of the fin strapping region is aligned with the inner edge of the via.

2.4 Summary

An understanding has been gained regarding the design optimization of FinFET source/drain and contact regions for LSTP and HP applications. Depending on the power/performance specification, the optimal source/drain design is fundamentally different, both in terms of material and geometry. Regarding LSTP design, DSS FinFETs are limited by the trade-off between I_{min} and R_c, which results in a small and shrinking SDE design space as L_G is scaled. In this power/performance regime, RSD FinFETs provide for both a larger design space and improved performance. However, for HP design, the increased C_{OV} in RSD FinFETs limits their utility to specific regimes of FP, H_{fin}, and silicide workfunction. Here, a specific flavor of DSS FinFETs, which uses a recessed strapping region, achieves a universal performance advantage over RSD FinFETs and conventional DSS FinFETs over all ranges of FP, H_{fin}, and silicide workfunction. This is achieved by what is effectively a dual sidewall spacer which decouples the flared silicide from the SDE regions (to reduce/eliminate silicide gating) while at the same time permitting lateral fin strapping and via contacts to the source/drain regions to reduce parasitic capacitance.

2.5 References

[1] 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.

[2] S. Zhu, J. Chen, M.-F. Li, S.J. Lee, J. Singh, C.X. Zhu, A. Du, C.H. Tung, A. Chin, D.L. Kwong, "N-Type Schottky Barrier Source/Drain MOSFET Using Ytterbium Silicide," *IEEE Elec. Dev. Lett.*, Vol .25, no. 8, 2004, pp. 565-567.

[3] 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.

[4] L.E. Calvet, H. Leubben, M.A. Reed, C. Wang, J.P. Snyder, J.R. Tucker, "Suppression of leakage current in Schottky barrier metal-oxide-semiconductor field-effect transistors," *J. App. Phys*, Vol. 91, no. 2, 2002, pp. 757-759.

[5] M. Fritze, C. L. Chen, S. Calawa, D. Yost, B. Wheeler, P. Wyatt, C.L. Keast, J. Snyder, J. Larson, "High-Speed Schottky-Barrier pMOSFET with f_T = 280 GHz," *IEEE Elec. Dev. Lett*., Vol. 25, no. 4, 2004, pp. 220-222.

[6] G. Larrieu, E. Dubois, "Schottky-Barrier Source/Drain MOSFETs on Ultrathin SOI Body With a Tungsten Midgap Gate," *IEEE Elec. Dev. Lett.*, Vol. 25, no. 12, 2004, pp. 801-803.

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

[8] R. A. Vega, "Schottky Field Effect Transistors and Schottky CMOS Circuitry," M.S. Thesis, Department of Microelectronic Engineering, Rochester Institute of Technology, Rochester, NY, 2006.

[9] T. Kinoshita, R. Hasumi, M. Hamaguchi, K. Miyashita, T. Komonda, A. Kinoshita, J. Koga, K. Adachi, Y. Toyoshima, T. Nakayama, S. Yamada, F. Matsuoka, "Ultra Low Voltage Operations in Bulk CMOS Logic Circuits with Dopant Segregated Schottky Source/Drain Transistors," *IEDM Tech. Dig.*, 2006, pp. 71-74.

[10] A. Kaneko, A. Yagashita, K. Yahashi, T. Kubota, M. Omura, K. Matsuo, I. Mizushima, K. Okano, H. Kawasaki, T. Izumida, T. Kanemura, N. Aoki, A. Kinoshita, J. Koga, S. Inaba, K. Ishimaru, Y. Toyoshima, H. Ishiuchi, K. Suguro, K. Eguchi, Y. Tsunashima, "High-Performance FinFET with Dopant-Segregated Schottky Source/Drain," *IEDM Tech. Dig.*, 2006, pp. 893-896.

[11] R. T. P. Lee, T.-Y. Liow, K.-M. Tan, A. E.-J. Lim, H.-S. Wong, P.-C. Lim, D. M.Y. Lai, G.-Q. Lo, C.-H. Tung, G. Samudra, D.-Z. Chi, Y.-C. Yeo, "Novel Nickel-Alloy Silicides for Source/Drain Contact Resistance Reduction in N-Channel Multiple-Gate Transistors with Sub-35nm Gate Length," *IEDM Tech. Dig.*, 2006, pp. 851-854.

[12] R. T. P. Lee, A. E.-J. Lim, K.-M. Tan, T.-Y. Liow, G.-Q. Lo, G. S. Samudra, D. Z. Chi, Y.-C. Yeo, "N-channel FinFETs with 25-nm Gate Length and Schottky-Barrier Source and Drain Featuring Ytterbium Silicide," *IEEE Elec. Dev. Lett.*, vol. 28, no. 2, pp. 164-167, Feb. 2007.

[13] 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.

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

[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] D. Connelly, C. Faulkner, P. A. Clifton, D. E. Grupp, "Fermi-level depinning for low-barrier Schottky source/drain transistors," *Appl. Phys. Lett.*, vol. 88, no. 1, 2006, 012105.

[17] 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.

[18] M. Tao, D. Udeshi, N. Basit, E. Maldonado, W.P. Kirk, "Removal of dangling bonds and surface states in silicon (001) with a monolayer of selenium," *App. Phys. Lett.*, Vol. 82, no. 10, 2003, pp. 1559-1561.

[19] M. Tao, S. Agarwal, D. Udeshi, N. Basit, E. Maldonado, W.P. Kirk, "Low Schottky barriers on *n*-type silicon (001)," *App. Phys. Lett.*, Vol. 83, no. 13, 2003, pp. 2593-2595.

[20] G. Song, M. Y. Ali, M. Tao, "A High Schottky-Barrier of 1.1 eV Between Al and S-Passivated p-Type Si(100) Surface," *IEEE Elec. Dev. Lett.*, vol. 28, no. 1, pp. 71-73, Jan. 2007.

[21] R. Saiz-Pardo, R. Perez, F. J. Garcia-Vidal, R. Whittle, F. Flores, "Systematic theoretical studies of the Schottky barrier control by passivating atomic intralayers," *Surface Science*, vol. 426, 1999, pp. 26-37.

[22] T. Yamauchi, A. Kinoshita, Y. Tsuchiya, J. Koga, K. Kato, "1 nm NiSi/Si Junction Design based on First-Principles Calculation for Ultimately Low Contact Resistance," *IEDM Tech. Dig.*, 2006, pp. 385-388.

[23] M. Zhang, J. Knoch, Q. T. Zhao, St. Lenk, U. Breuer, S. Mantl, "Schottky barrier height modulation using dopant segregation in Schottky-barrier SOI-MOSFETs," *Proc. ESSDERC*, 2005, pp. 457-460.

[24] C. H. Ko, H. W. Chen, T. J. Wang, T. M. Kuan, J. W. Hsu, C. Y . Huang, C. H. Ge, L.S. Lai, W. C. Lee, "NiSi Schottky Barrier Process-Strained Si (SB-PSS) CMOS Technology for High Performance Applications," *VLSI Tech Dig.*, 2006, pp. 80-81.

[25] L. K. Bera, Y. F. Lim, S. J. Tan, W. Y. Loh, B. Ramana Murthy, N. Singh, Y. Rong, C. H. Tung, H. S. Nguyen, R. Kumar, G. Q. Lo, N. Balsubramanian, D. L. Kwong, "Dopant-Segregated Ni-Silicide Schottky-Source/Drain CMOS on Starined-Si/SiGe Multiple Quantum-Well Channel on Bulk-Si," *Proc. ESSDERC*, 2006, pp. 290-293.

[26] Y. Nishi, A. Kinoshita, D. Hagishima, J. Koga, "Experimental Study on Performance Improvement in Dopant-Segregated Schottky Metal-Oxide-Semiconductor Field-Effect Transistors," *Jpn. J. Appl. Phys.*, vol. 47, no. 1, pp. 99-103, 2008.

[27] A. Kinoshita, T. Kinoshita, Y. Nishi, K. Uchida, S. Toriyama, R. Hasumi, J. Koga, "Comprehensive Study of Injection Velocity Enhancement in Dopant-Segregated Schottky MOSFETs," *IEDM Tech. Dig.*, pp. 79-82, 2006.

[28] H. Onoda, K.Miyashita, T. Nakayama, T. Kinoshita, H. Nishimura, A. Azuma, S. Yamada, F. Matsuoka, "0.7 V SRAM Technology with Stress-Enhanced Dopant Segregated Schottky (DSS) Source/Drain Transistors for 32 nm Node," *VLSI Tech. Dig.*, pp. 76-77, 2007.

[29] H.-W. Chen, C.-H. Ko, T.-J. Wang, C.-H. Ge, K. Wu, W.-C. Lee, "Enhanced Performance of Strained CMOSFETs Using Metallized Source/Drain Extension (M-SDE)," *VLSI Tech. Dig.*, pp. 118-119, 2007.

[30] A. Kaneko, A. Yagashita, K. Yahashi, T. Kubota, M. Omura, K. Matsuo, I. Mizushima, K. Okano, H. Kawasaki, T. Izumida, T. Kanemura, N. Aoki, A. Kinoshita, J. Koga, S. Inaba, K. Ishimaru, Y. Toyoshima, H. Ishiuchi, K. Suguro, K. Eguchi, Y. Tsunashima, "High-Performance FinFET with Dopant-Segregated Schottky Source/Drain," *IEDM Tech. Dig.*, pp. 893-896, 2006.

[31] 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, pp. 8040-8046, Jun. 1997.

[32] 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, pp. 8047-8051, Jun. 1997.

[33] Z. Zhang, J. Lu, Z. Qiu, P.-E. Hellström, M. Ostling, S.-L. Zhang, "Performance Fluctuation of FinFETs With Schottky Barrier Source/Drain," *IEEE Elec. Dev. Lett.*, vol. 29, no. 5, pp. 506-508, May 2008.

[34] *User's Manual for Sentaurus Device*, Synopsys Co., Mountainview, CA.

[35] M. Jang, Y. Kim, M. Jeon, C. Choi, I. Baek, S. Lee, B. Park, "N_2-Annealing Effects on Characteristics of Schottky-Barrier MOSFETS," *IEEE Trans. Elec. Dev.*, vol. 53, no. 8, pp. 1821-1825, Aug. 2006.

[36] M. H. Unewisse, J. W. V. Storey, "Conduction mechanisms in erbium silicide Schottky diodes," *J. Appl. Phys.*, vol. 73, no. 8, pp. 3873-3879, Apr. 1993.

[37] J. A. Knapp, S. T. Picraux, C. S. Wu, S. S. Lau, "Kinetics and morphology of erbium silicide formation," *J. Appl. Phys.*, vol. 58, no. 10, pp. 3747-3757, Nov. 1985.

[38] P. Muret, T. A. Nguyen Tan, N. Frangis, J. Van Landuyt, "Unpinning of the Fermi level at erbium silicide/silicon interfaces," *Phys. Rev. B*, vol. 56, no. 15, pp. 9286-9289, Oct. 1997.

[39] M. Q. Huda, K. Sakamoto, "User of $ErSi_2$ in source/drain contacts of ultra-thin SOI MOSFETs," *Materials Science and Engineering B*, vol. 89, 2002, pp. 378-381.

[40] E. J. Tan, K. L. Pey, D. Z. Chi, P. S. Lee, L. J. Tang, "Improved Electrical Performance of Erbium Silicide Schottky Diodes Formed by Pre-RTA Amorphization of Si," *IEEE Elec. Dev. Lett.*, vol. 27, no. 2, pp. 93-95, Feb. 2006.

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

[42] K. Shenai, E. Sangiorgi, R. M. Swanson, K. C. Saraswat, R. W. Dutton, "Modeling and Characterization of Dopant Redistributions in Metal and Silicide Contacts," *IEEE Trans. Elec. Dev.*, vol. 32, no. 4, pp. 793-799, Apr. 1985.

[43] R. A. Vega, "Comparison Study of Tunneling Models for Schottky Field Effect Transistors and the Effect of Schottky Barrier Lowering," *IEEE Trans. Elec. Dev.*, vol. 53, no. 7, pp. 1593-1600, July 2006.

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

[45] R. Rengel, E. Pascual, M. J. Martin, "Injected Current and Quantum Transmission Coefficient in Low Schottky Barriers: WKB and Airy Approaches," *IEEE Elec. Dev. Lett.*, vol. 28, no. 2, pp. 171-173, Feb. 2007.

[46] J. Appenzeller, M. Radosavljevic, J. Knoch, Ph. Avouris, "Tunneling Versus Thermionic Emission in One-Dimensional Semiconductors," *Phys. Rev. Lett.*, vol. 92, no. 4, 048301, Jan. 2004.

[47] A. Kinoshita, T. Kinoshita, Y. Nishi, K. Uchida, S. Toriyama, R. Hasumi, J. Koga, "Comprehensive Study on Injection Velocity Enhancement in Dopant-Segregated Schottky MOSFETs," *IEDM Tech. Dig.*, 2006, pp. 79-82.

[48] A. Benedetti, H. Bender, C. Torregiani, M. Van Dal, K. Maex, "Nanometer scale characterization of $CoSi_2$ and NiSi induced strain in Si by convergent beam electron diffraction," *Materials Science and Engineering B*, vol. 114-115, pp. 61-66, 2004.

[49] M. Tsuchiaki, K. Ohuchi, A. Nishiyama, "Suppression of Thermally Induced Leakage of NiSi-Silicided Shallow Junctions by Pre-Silicide Fluorine Implantation," *Jpn. J. Appl. Phys.*, vol. 44, no. 4A, pp. 1673-1681, 2005.

[50] 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," *IEDM Tech. Dig.*, 2004, pp. 168-169.

[51] R. F. Pierret, *Semiconductor Device Fundamentals*. Reading, MA: Addison-Wesley, 1996, pp. 483-486.

[52] A. Kinoshita, C. Tanaka, K. Uchida, J. Koga, "High-performance 50-nm-Gate-Length Schottky-Source/Drain MOSFETs with Dopant-Segregation Junctions," *VLSI Tech. Dig.*, 2005, pp. 158-159.

[53] H.-W. Chen, C.-H. Ko, T.-J. Wang, C.-H. Ge, K. Wu, W.-C. Lee, "Enhanced Performance of Strained CMOSFETs Using Metallized Source/Drain Extension (M-SDE)," *VLSI Tech. Dig.*, 2007, pp. 118-119.

[54] Y.-J. Zhai, J.-F. Kang, G. Du, R.-Q. Han, X.-Y. Liu, "An Assessment of the Performance for Double Gate Schottky Barrier MOSFETs with Modulated Back Gate," *8th International Conference on Solid-State and Integrated Circuit Technology*, pp. 87-89, Oct. 2006.

[55] H. Zhao, Y.-C. Yeo, S. C. Rustagi, G. S. Samudra, "Analysis of the Effects of Fringing Electrid Field on FinFET Device Performance and Structural Optimization Using 3-D Simulation," *IEEE Trans. Elec. Dev.*, vol. 55, no. 5, pp. 1177-1184, May 2008.

[56] Z. Qiu, Z. Zhang, M. Ostling, S.-L. Zhang, "A Comparative Study of Two Different Schemes to Dopant Segregation at NiSi/Si and PtSi/Si Interfaces for Schottky Barrier Height Lowering," *IEEE Trans. Elec. Dev.*, vol. 55, no. 1, pp. 396-403, Jah. 2008.

[57] H.-S. Wong, L. Chan, G. Samudra, Y.-C. Yeo, "Sub-0.1-eV Effective Schottky-Barrier Height for NiSi on n-Type Si (100) Using Antimony Segregation," *IEEE Elec. Dev. Lett.*, vol. 28, no. 8, pp. 703-705, Aug. 2007.

[58] J. Yuan, J. C. S. Woo, "Tunable Workfunction in Fully Nickel-Silicided Polysilicon Gates for Metal Gate MOSFET Applications," *IEEE Elec. Dev. Lett.*, vol. 6, no. 2, pp. 87-89, Feb. 2005.

[59] E. Bucher, S. Schulz, M. Ch. Lux-Steiner, P. Munz, U. Gubler, F. Greuter, "Work Function and Barrier Heights of Transition Metal Silicides," *Appl. Phys. A.*, vol. A 40, pp. 71-77, 1986.

[60] K. Sano, M. Hino, M. Ooishi, K. Shibahara, "Workfunction Tuning Using Various Impurities for Fully Silicided NiSi Gate," *Jpn. J. Appl. Phys.*, vol. 44, no. 6A, pp. 3774-3777, 2005.

[61] Y. Nishi, Y. Tsuchiya, A. Kinoshita, A. Hokazano, J. Koga, "Successful Enhancement of Metal Segregation at NiSi/Si Junction Through Pre-amorphization Technique," *VLSI Tech. Dig.*, 2008, pp. 192-193.

[62] M. Nawaz, S. Decker, L.-F. Giles, W. Molzer, T. Schulz, K. Schrufer, R. Mahnkopf, "Device Design Evaluation of Multigate FETs Using Full 3D Process and Device TCAD Simulation," *Simulation of Semiconductor Processes and Devices*, vol. 12, pp. 401-404, Sept. 2007.

[63] R. Duffy, G. Curatola, B. J. Pawlak, G. Doornbos, K. van der Tak, P. Breimer, J. G. M. van Berkum, F. Roozeboom, "Doping fin field-effect transistor sidewalls: Impurity dose retention in silicon due to high angle incident ion implants and the impact on device performance," *J. Vac. Sci. Technol. B*, vol. 26, no. 1, pp. 402-407, Jan/Feb 2008.

[64] K. Kobayashi, K. Okuyama, H. Sunami, "Plasma doping induced damages associated with source/drain formation in three-dimensional beam-channel MOS transistor," *Microelectronic Engineering*, vol. 84, pp. 1631-1634, 2007.

[65] B. Mizuno, Y.Sasaki, "Aiming for The Best Matching between Ultra-Shallow Doping and Milli- to Femto-Second Activation," *IEEE Advanced Thermal Processing of Semiconductors*, pp. 1-10, 2007.

[66] C. M. Ransom, T. N. Jackson, J. F. DeGelormo, C. Zeller, D. E. Kotecki, C. Graimann, D. K. Sadana, J. Benedict, "Shallow n+ Junctions in Silicon by Arsenic Gas-Phase Doping," *J. Electrochem. Soc.*, vol. 141, no. 5, pp. 1378-1381, May 1994.

[67] J. C. Ho, R. Yerushalmi, Z. A. Jacobson, Z. Fan, R. L. Alley, A. Javey, "Controlled nanoscale doping of semiconductors via molecular monolayers," *Nature Materials*, vol. 7, pp. 62-67, Jan. 2008.

[68] J. Kedzierski, M. Ieong, E. Nowak, T. S. Kanarsky, Y. Zhang, R. Roy, D. Boyd, D. Fried, H.-S. P. Wong, "Extension and Source/Drain Design for High-Performance FinFET Devices," *IEEE Trans. Elec. Dev.*, vol. 50, no. 4, pp. 952-958, Apr. 2003.

[69] H. Shang, L. Chang, X. Wang, M. Rooks, Y. Zhang, B. To, K. Babich, G. Totir, Y. Sun, E. Kiewra, M. Ieong, W. Haensch, "Investigation of FinFET devices for 32nm technologies and beyond," *VLSI Tech. Dig.*, pp. 54-55, 2006.

[70] R. A. Vega, T.-J. King Liu, "A Comparative Study of Dopant-Segregated Schottky and Raised Source/Drain Double-Gate MOSFETs," *IEEE Trans. Elec. Dev.*, vol. 55, no. 10, pp. 2665-2677, Oct. 2008.

[71] M. Guillorn, J. Chang, A. Bryant, N. Fuller, O. Dokumaci, X. Wang, J. Newbury, K. Babich, J. Ott, B. Haran, R. Yu, C. Lavoie, D. Klaus, Y. Zhang, E. Sikorski, W. Graham, B. To, M. Lofaro, J.Tornello, D. Koli, B. Yang, A. Pyzyna, D. Neumeyer, M. Khater, A. Yagashita, H. Kawasaki, W. Haensch, "FinFET Performance Advantage at 22nm: An AC perspective," *VLSI Tech. Dig.*, pp. 12-13, 2007.

[72] S.-P. Jeng, K. Taylor, T. Seha, M.-C. Chang, J. Fattaruso, R. H. Havemann, "Highly Porous Interlayer Dielectric For Interconnect Capacitance Reduction," *VLSI Tech. Dig.*, pp. 61-61, 1995.

Chapter 3

Sub-10 nm Double Gate MOSFET Design

3.1 Introduction

The fundamental limit for MOSFET channel-length scaling is thought to be set by direct source-to-drain tunneling (DSDT), when the source potential barrier becomes so narrow that tunneling straight through this barrier dominates the subthreshold leakage. Although DSDT has been experimentally demonstrated [1] and some nanoscale modeling efforts have considered DSDT [2]–[5], little work [4] has focused on engineering the shape of the DSDT barrier to minimize DSDT. Also, DSDT modeling studies to date have largely ignored fringing-field effects which would be very significant in a practical device. This work aims to illustrate and quantify the benefits of various electrostatic, doping, and geometric design approaches that may be employed to minimize DSDT and maximize electrostatic integrity, as well as to point out the implications for gate pitch (GP) and fin pitch (FP) scaling. Of particular interest in this study are the roles of source/drain junction abruptness, Schottky barrier height (SBH), and gate stack design in symmetric double-gate (SDG) dopant-segregated Schottky (DSS) source/drain MOSFETs in the DSDT regime. Only low operating power (LOP) specifications are considered for this study, since leakage current is less of a concern for high performance (HP) devices. Low standby power (LSTP) devices, while requiring even lower off-state leakage than LOP devices, are not considered in this study because it was shown in [7] that the DSS source/drain design is not optimal for LSTP applications; rather, the raised source/drain structure is optimal. This study begins with a relatively simple SDG structure, to which refinements are made in the course of the paper as the importance of various fringing-field effects for design optimization in the DSDT regime are elucidated.

3.2 Modeling Approach

The SDG DSS MOSFET in Fig. 3.1 is initially considered, with gate length L_G = 3 nm. The gate oxide thickness t_{ox} = 1 nm and the body thickness t_{body} = 3 nm, which is assumed to be a lower limit in consideration of quantization effects and practical manufacturing limitations. Only a silicon body is considered in this study (with a p-type doping of 1×10^{15} cm^{-3}), as elsewhere [4], [5] it was shown that germanium and III-IV body regions offer no performance

advantage in the DSDT regime due to their lower carrier tunneling masses. The gate sidewall spacers have a length L_{sp} and their dielectric constant ε_{spacer} is varied to investigate the influence of gate fringing fields on DSDT. The metal gate height t_{gate} is varied for the same reason.

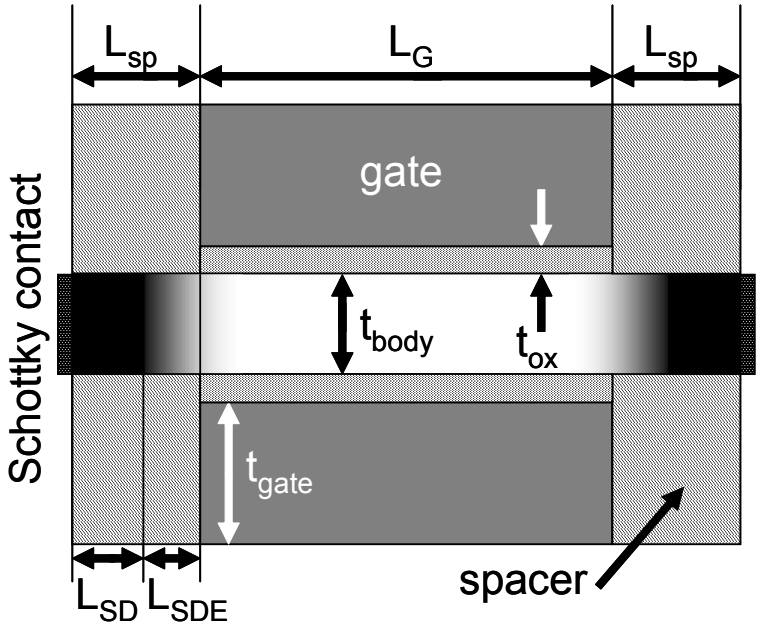

Fig. 3.1. Schematic cross-section of the SDG DSS MOSFET initially modeled in this study.

The source/drain extension (SDE) consists of two regions. First is a constant doping region of length L_{SD} and concentration N_{SDE}, extending from the Schottky barrier (SB) contact inwards toward the gate. Second is a graded (Gaussian) doping region with peak concentration N_{SDE} and length L_{SDE}, where L_{SDE} is the distance from the peak position to where the doping concentration drops by 1 decade. Here, $N_{SDE} = 1\text{x}10^{20}$ cm^{-3} and the effective gate length L_{eff} is defined by the distance between the source and drain profiles at a concentration of $1\text{x}10^{19}$ cm^{-3}. Although Gaussian profiles do not have a constant gradient, here the abruptness is estimated simply as L_{SDE} (e.g., L_{SDE} = 5 nm means the source/drain junction abruptness = 5 nm/dec.). Here, $L_{sp} = L_{SD} + L_{SDE}$ and $L_{eff} = L_G$ unless otherwise noted. All dopant profiles are treated as continuum profiles, meaning that random dopant fluctuation (RDF) is ignored here. At small enough scales, RDF gives rise to hot spots at the contact interface where the SB is smaller and more current is injected. As will be shown in the next chapter, the effect of RDF on contact resistance will not present a significant barrier to CMOS scaling. Thus, continuum mode doping is used here to model the nominal structure with no RDF-induced variations. The silicide gating effect [8] is initially ignored, although it is considered later to demonstrate how its exclusion can lead to erroneous conclusions with regard to device performance and design optimization. For LOP applications, the off-state current I_{OFF} = 24 nA/μm and V_{DD} = 0.45 V at the end of the CMOS technology roadmap [9].

Ideal V_t tuning through gate workfunction engineering is assumed (to achieve I_{OFF} = 24 nA/μm at V_{GS} = 0 V and $V_{DS} = V_{DD}$) and bandgap narrowing due to heavy doping is included (as this will lower the SBH slightly [7]). Gate leakage is ignored. A conventional drift-diffusion transport model with dopant- and field-dependent mobility is used to model thermal current. Although the L_G values used here are small and several modeling studies have assumed ballistic transport in this regime with I_{ON} exceeding 2 mA/μm [4], [7], [10], empirical evidence [11], [12] is not so

optimistic, even when significant amounts of strain are applied [13]. The mobility is treated here as independent of the gate dielectric material. The SB contacts are assumed to have abrupt potential profiles and, although the SBH is varied in this study, SB lowering (SBL) is ignored, due to the limitations of the TCAD software. (This is acceptable, since the optimal structure ends up having SBH = 0 eV and, for SBH > 0 eV, the lateral field at the SB junction is already too high, due to the SDE regions, for the gate electrode to have any significant further effect.) Longitudinal tunneling current (*i.e.*, SB tunneling and DSDT) is modeled using a non-local 1-D Schrödinger solution with electron and hole effective tunneling masses (effective Richardson's constants) of $0.19*m_0$ (112 A/cm^2*K^2) and $0.16*m_0$ (32 A/cm^2*K^2), respectively. Thermal current over the SB is modeled with the "Schottky" contact model in Sentaurus Device. Band-to-band tunneling (BTBT) is ignored here, due to convergence issues with the combined use of the BTBT models and the Schrödinger tunneling model. This is acceptable, though, since the low V_{DD} would reduce BTBT current to far below the LOP I_{OFF} specification [7].

Quantization effects in the transverse direction are included in I-V simulations through the use of a non-local 1-D Schrödinger quantization model. This model accounts for both the charge redistribution within the body region as well as subband splitting due to quantization. (Anisotropic electron effective masses of $0.19*m_0$ and $0.92*m_0$ are used here for electrons, while the warped band structure is used for holes, as described in [6] and [10].) This is different from density gradient method (DGM) formalisms [6], which only account for charge redistribution through the addition of a potential-dependent effective band edge shift in the carrier density computation. For the C-V simulations here, it is only the charge distribution that matters, in which case the DGM formalism is used to improve simulation speed and convergence.

Admittedly, the TCAD software does not yet offer a means of extracting the individual subband contributions to DSDT, so only one effective tunneling mass (listed above for electrons and holes) can be applied to the Schrödinger tunneling model when calculating DSDT and tunneling from the SB contact into the various subbands. This is acceptable, though, since the transport (longitudinal) direction is dominated by the light subbands (4 out of 6 of them are light with $0.19*m_0$), which are heavy ($0.92*m_0$) in the quantization (transverse) direction. Thus, the electrons with the highest tunneling probability (due to the light longitudinal mass) also have the highest population due to their lower energy (heavy transverse mass, less quantization) and so are expected to dominate the I-V behavior. For the case of t_{body} = 3 nm, the modeled band edge shifts (and therefore the SBH increases) for the transverse light and heavy subbands are 122 meV and 37 meV, respectively. For simplicity, the SBH values from this point forward are set to what they would be in a non-quantized system, and so these offsets must be taken into account by the reader. For example, if the SBH is set to 0.2 eV, then the actual SBH would be 0.322 eV and 0.237 eV to the longitudinal heavy and light subbands, respectively (neglecting bandgap narrowing due to heavy doping). Finally, for all conduction band profiles shown, quantization is excluded and values are taken from the center of the body region. This is done to simplify what is otherwise a very complex extraction in a data field with orthogonal non-local grids, but at no cost to its illustrative power.

3.3 Effect of SDE Junction Abruptness

Fig. 3.2(a) shows *SS* versus L_{SDE} for L_G = 3 nm, with and without tunneling. Since DSDT adds to the subthreshold current which is normally only thermal current, an increase in DSDT will

result in an increase in *SS* [1]-[3]. Conventionally, L_{SDE} is scaled with L_G to keep short channel effects (SCE) at bay; however, in the DSDT regime, as Fig. 3.2(a) shows for $L_{SDE} < 7$ nm, if the SDE profile is too sharp for the same L_{eff}, *SS* will increase significantly. This is because the off-state energy band profile becomes narrower as L_{SDE} is reduced (Fig. 3.2(b)), resulting in higher DSDT. By broadening out the SDE profile, the average SDE doping under the spacer drops, thus broadening the off-state energy band profile, reducing DSDT at low electron energies and, consequently, *SS*. The result is simply an increase in L_{elec}, which is different from L_{eff} in that L_{eff} is defined at constant doping while L_{elec} is defined at constant energy. So, although L_{eff} is the same for all cases in Fig. 3.2(b), L_{elec} increases with L_{SDE}, thus improving gate control. In the ITRS [9], only short channel effects are considered in determining L_{SDE} at each technology node. The formulism used assumes a 3-decade drop in concentration over the lateral extent of the junction, which itself is 60% of the vertical junction, which is defined as $0.5*L_G$. This results in $L_{SDE} = 0.1*L_G$ at each technology node. Although this formulism is not optimal in the DSDT regime, the solution is not as simple as to make L_{SDE} as large as possible, since second-order effects will influence the optimal L_{SDE} value. This is covered in more detail in Section 3.7, which discusses gate-underlapped structures ($L_G < L_{eff}$) and how silicide gating affects their optimization; for now, the analysis here continues with $L_G = L_{eff}$.

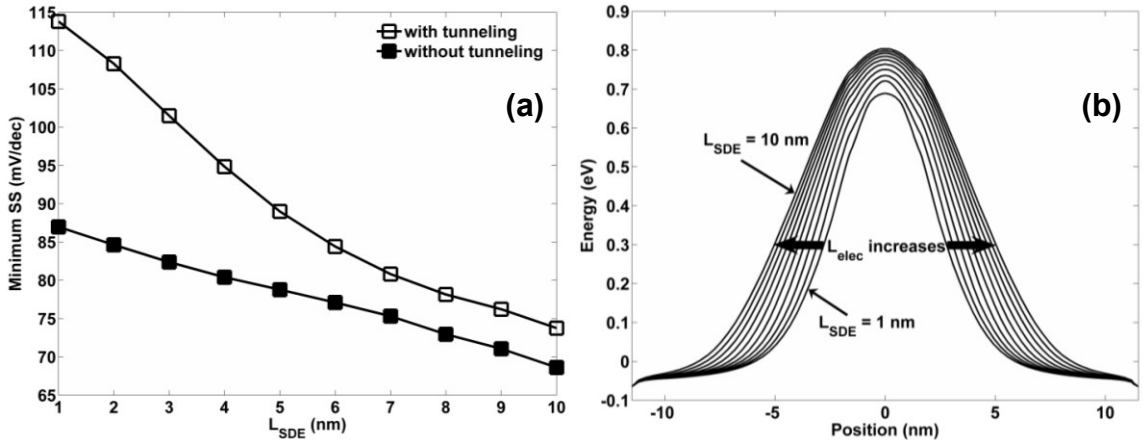

Fig. 3.2. (a) Minimum *SS* vs. L_{SDE} with and without tunneling and (b) Off-state lateral conduction band profile at the center of the body region for different SDE junction abruptness values ($\Delta L_{SDE} = 1$ nm). $L_{sp} = 10$ nm, $t_{gate} = 6$ nm, SBH = 0 eV, and the sidewall spacer is silicon nitride. In (a) $V_{DS} = 50$ mV, to minimize the contribution of short channel effects and emphasize the contribution of DSDT. In (b), $V_{GS} = V_{DS} = 0$ V, $L_{sp} = 10$ nm, $t_{gate} = 6$ nm, and SBH = 0 eV. The sidewall spacer is silicon nitride and the gate workfunction is set to 5.2 eV.

3.4 Effect of Schottky Barrier Height

As one would expect, *SS* increases as L_{sp} is scaled down. This is because L_{sp} places an upper limit on L_{SDE} for constant L_{eff}, thereby placing an upper limit on L_{elec} and therefore a lower limit on DSDT. Although the simple fix is to keep L_{sp} large to allow for high L_{SDE} values for constant L_{eff}, this limits the ability to scale the GP and therefore device density. One method for keeping the off-state energy barrier wide at low electron energies while scaling L_{sp} is to increase the SBH of the contacts. This will result in a SB depletion region extending from the SB contact outward toward the channel. If the contact and the gate edge are in close enough proximity, the SB depletion region and the gate induced depletion region will overlap, effectively increasing the

DSDT barrier width. An example of this is shown in Fig. 3.3.

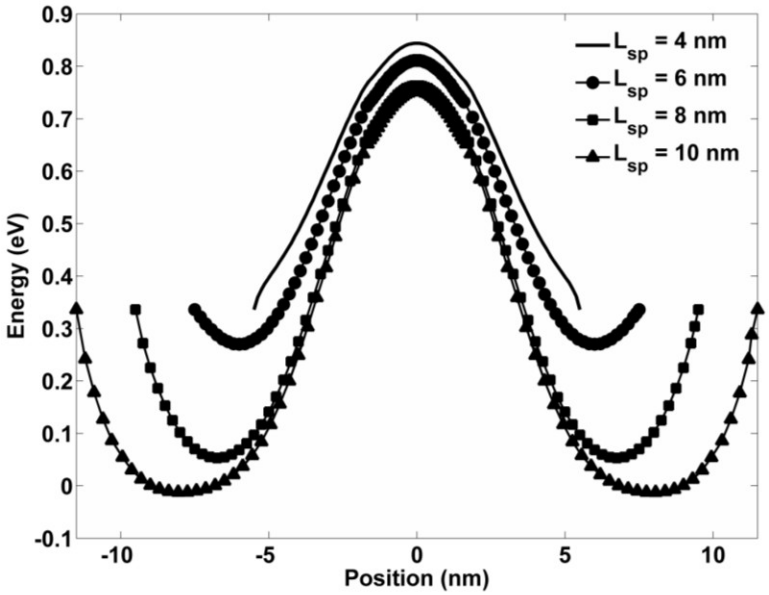

Fig. 3.3. Off-state conduction band profile at the center of the body region for different SDE junction abruptness values. $V_{GS} = V_{DS} = 0$ V, $L_{SDE} = 4$ nm, $L_{sp} = 10$, 8, 6, and 4 nm, $t_{gate} = 6$ nm, and SBH = 0.4 eV. The sidewall spacer is silicon nitride and the gate workfunction is set to 5.2 eV.

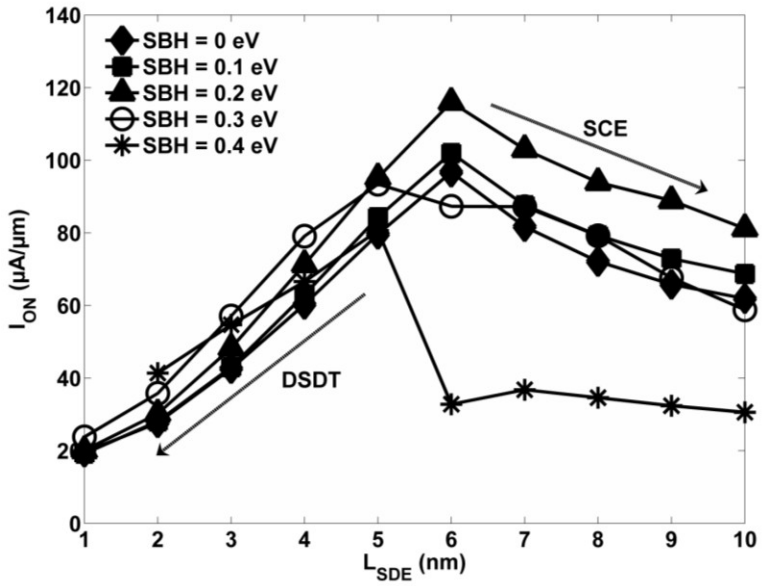

Fig. 3.4. I_{ON} vs. L_{SDE} for different SBH values. $L_{sp} = 6$ nm, $t_{gate} = 6$ nm, and the sidewall spacer is silicon nitride.

This is perhaps the most effective use of finite-SBH contacts in nanoscale MOSFETs, as otherwise they are detrimental to contact resistance. For L_G large enough such that DSDT is negligible, the SBH should be as small as possible (ideally zero) to keep *SS* low and I_{ON} as large as possible. In the DSDT regime, however, SBH tuning can be used as one of several techniques to enable L_G and device pitch scaling while maintaining electrostatic integrity. On the other hand, there are several trade-offs imposed by increasing the SBH in this regime, the most important of which is an increase in contact resistance R_c and a resulting decrease in I_{ON}. (Others

include an increase an ambipolar leakage and the challenge of finely tuning the SBH while keeping N_{SDE} relatively high, since heavy doping reduces the SBH to zero or near-zero values through image force, dipole, and bandgap narrowing effects [7], [14].) Thus, if the SBH is too high, I_{ON} will drop due to R_{c_2} and if the SBH is too low, I_{ON} will drop due to a DSDT-induced increase in *SS*, for small L_{sp}. Fig. 3.4 shows an example for L_G = 3 nm and L_{sp} = 6 nm. In this case, the maximum I_{ON} is achieved with SBH = 0.2 eV and L_{SDE} = 6 nm. For L_{SDE} > 6 nm, I_{ON} drops because L_{eff} drops due to $L_{SDE} > L_{sp}$ (in which case the peak of the SDE Gaussian profile is held at the SB contact), meaning that SCE degrades *SS*. Otherwise, I_{ON} increases with L_{SDE} in most cases for L_{SDE} < 6 nm, due to the effect of the SBH on reducing DSDT and therefore improving *SS*. For SBH = 0.3 eV and 0.4 eV, I_{ON} actually drops sooner, for L_{SDE} > 5 nm. This is because as L_{SDE} increases, the average doping under the sidewall spacer decreases. This increases the SB width and reduces I_{ON} due to lower SB tunneling current.

3.5 Effect of Gate Sidewall Spacers

An alternative approach for increasing L_{elec} is to leverage the fringing fields extending from the gate sidewall to deplete the SDE region in the off state. (An additional advantage is a reduction in SDE and contact resistance in the on-state due to majority carrier accumulation). Up to this point, a constant t_{gate} = 6 nm was assumed; however, increasing t_{gate} will increase the sidewall fringing fields and therefore L_{elec}. Eventually this effect will saturate due to the $1/t_{gate}^2$ dependence of electric field strength, and so increasing this effect further will require more efficient gate sidewall coupling to the SDE regions, *i.e.* increasing ε_{spacer}. This has been proposed before, for Schottky barrier (SB) MOSFETs [15] and conventional source/drain MOSFETs using planar [16] and FinFET [17] architectures. In addition to nominal performance improvements with high-k sidewall spacers, a compelling case is made for reduced susceptibility to sources of performance variation [17].

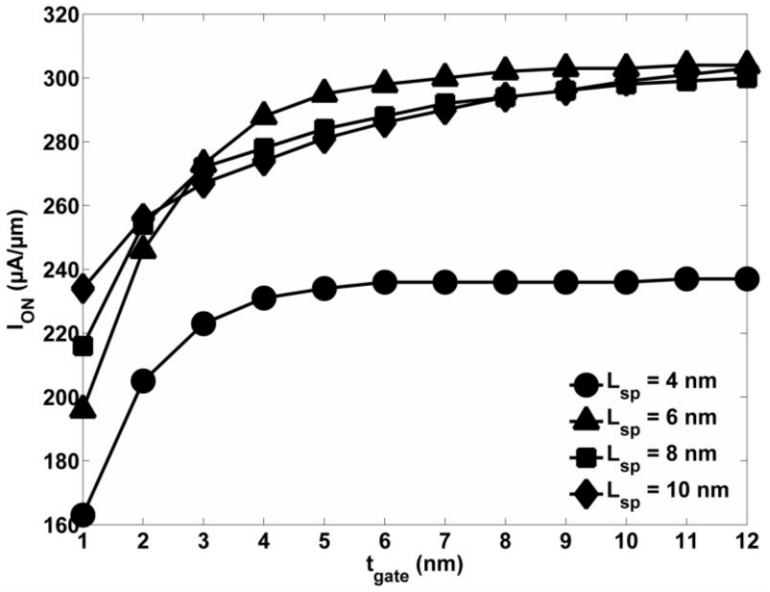

Fig. 3.5. I_{ON} vs. t_{gate} for $L_{sp} = L_{SDE}$ = 10, 8, 6, and 4 nm, and SBH = 0 eV, for HfO_2 sidewall spacers (ε_{spacer} = 23).

Fig. 3.5 demonstrates that there exists an optimum L_{sp} which balances the trade-off between series resistance R_{ser} and DSDT. For large L_{sp}, R_{ser} limits I_{ON}, while for small L_{sp}, DSDT increases *SS* and sharply reduces I_{ON}, due to the reduction in the maximum possible L_{elec}. (The L_{elec} limit is not defined by the spacers but instead the distance between the metallic contacts, although here it is assumed that there is no lateral silicidation under the spacer when the contacts are formed.) SCE also contributes to the I_{ON} reduction, but its effect is smaller in this case. Although not shown here, with tunneling excluded from the I-V simulation, I_{ON} drops from 289 μA/μm to 266 μA/μm as L_{sp} is reduced from 6 nm to 4 nm, with t_{gate} = 12 nm. With tunneling included, a larger I_{ON} drop, as shown in Fig. 3.5, from 304 μA/μm to 237 μA/μm, takes place under the same conditions. This is a recurring theme throughout the course of this chapter, that structural changes which affect DSDT also affect SCE in the same manner due to the similar dependence of each on L_{elec}. To quantify the relative DSDT and SCE contributions experimentally, the critical temperature T_{crit} below which *SS* remains constant and DSDT determines I_{OFF} should be evaluated as a function t_{gate}, L_{sp}, source/drain anneal conditions, etc. [1].

As Fig. 3.5 shows, $t_{gate} \sim L_{sp}$ is a reasonable design rule for part of this multi-dimensional optimization. Additionally, Figs. 3.2(a) and 3.4 suggest that $L_{sp} \sim L_{SDE}$ gives the optimum performance almost irrespective of SBH, and so the two remaining variables to optimize are SBH and ε_{spacer}. However, as Fig. 3.6 shows, increasing SBH does not correlate to improved I_{ON} when high-k spacers are employed. For $\varepsilon_{spacer} < 8$, increasing the SBH compensates for the low gate sidewall coupling to the SDE regions in the off state. As ε_{spacer} increases, the gate fringing fields alone are sufficient to turn the device off. Even for small L_{sp}, where SBH increase would be the most useful, the effect of ε_{spacer} on improving electrostatic integrity is clearly much stronger.

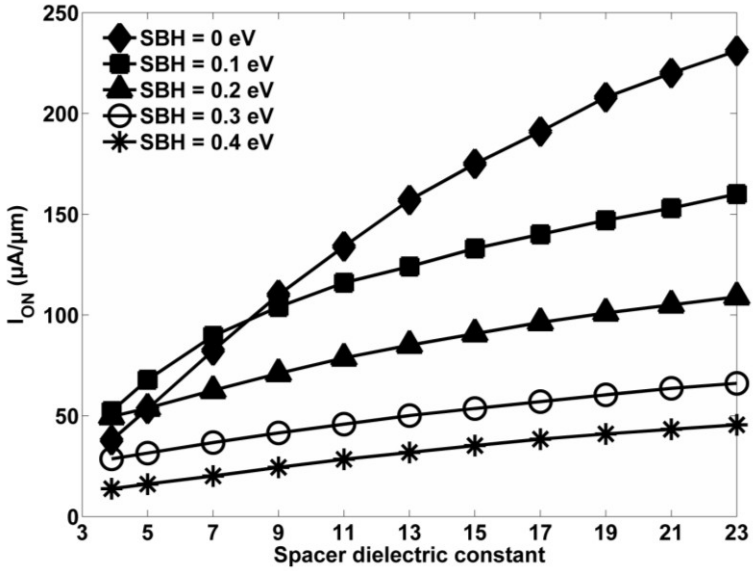

Fig. 3.6. I_{ON} vs. ε_{spacer} for $L_{sp} = L_{SDE} = t_{gate}$ = 4 nm and varying SBH.

The implication here is that L_G no longer becomes the primary component in limiting GP, which becomes determined increasingly by L_{sp} (to maximize L_{elec}). Shrinking L_{sp} necessitates an increase in ε_{spacer} to maintain electrostatic integrity, but this will increase the parasitic fringing capacitance from the gate sidewall to the contact vias and flared source/drain regions. However,

neglecting capacitance for now and simplistically assuming the optimal L_{sp} as giving the highest I_{ON}, then L_{sp} = 6 nm (high-k) for L_G = 3 nm, as Fig. 3.5 shows. Going much further and assuming an ideal overlay error of zero and that both the STI and source/drain via lengths can scale with L_G (i.e., they are both 3 nm), and also assuming that either end of the source/drain via directly contacts the edge of the STI region and sidewall spacer, one ends up with an ultimately small GP of 24 nm (GP = $L_G + 2*L_{sp} + 2*L_{via} + L_{STI}$). One option for reducing GP further is to eliminate the STI regions between devices sharing a node within a circuit by using the source/drain silicide as self-aligned inter-device isolation, provided SOI substrates are used to prevent cross-talk, so that GP = $L_G + 2*L_{sp} + L_{via}$. This would reduce the GP to 18 nm, which is effectively one technology generation. This approach has been demonstrated empirically in [18] and is a promising complementary approach to high-k spacers for GP scaling. One might also consider forming a gate-underlapped structure, such that $L_{eff} > L_G$, to enable L_{sp} and therefore GP scaling. However, the increase in L_{eff} and therefore L_{elec} by scaling L_{SDE} (with L_{SD} = 0, which would give the largest L_{eff} increase) is more than offset by the reduction in maximum possible L_{elec} when L_{sp} is reduced. As a result, I_{ON} drops due to DSDT and SCE. This is covered later in Section 3.7, but the important point here is that there is a lower limit to L_{sp} and it is determined in part by DSDT.

3.6 Effect of Gate Dielectric

High-k gate dielectrics have received much attention for scaling equivalent oxide thickness (EOT) [19]-[21] and are currently in production with metal gate technology at the 45 nm node [21]. Although the purpose of high-k is to scale EOT while keeping gate leakage low, this results in the bottom of the gate electrode being placed farther away from the body region. As a result, the effect of gate fringing fields on reducing DSDT is diminished. This, in addition to the already well-known fringe-induced barrier lowering (FIBL) effect [22], will increase *SS* and reduce I_{ON}. This effect can be countered somewhat by increasing ε_{spacer} or by reducing EOT and therefore the physical gate dielectric thickness t_{gd} at constant ε_{gd} (gate dielectric constant). Ultimately, though, there exists a limit to electrostatic control regardless of ε_{spacer}, as the bottom of the gate electrode is pulled farther away from the body region due to the increased ε_{gd} (Fig. 3.7).

It is worth noting that, for the vertical double-gate FinFET structure, t_{gd} will limit FP scaling unless EOT is scaled down, since in this regime t_{gd} is an appreciable portion of FP. For example, with t_{body} = 3 nm, t_{gate} = 6 nm, and t_{gd} = 5.9 nm (1 nm EOT with HfO_2 gate dielectric), FP = $t_{body} + 2*(t_{gate} + t_{gd})$ = 26.8 nm. Here, $2*t_{gd}$ contributes to 44% of FP. Reducing EOT to 0.5 nm reduces FP to 20.9 nm and the $2*t_{gd}$ contribution to 28% of FP. Scaling FP further requires scaling t_{gate}, which in this example accounts for 57% of FP. However, t_{gate} scaling below a certain height ($L_{sp} - t_{gd}$ + 1 nm) will reduce I_{ON}, as shown in Fig. 3.5 for $t_{gd} = t_{ox}$ = 1 nm. For EOT = 0.5 nm and ε_{gd} = 23, this means that t_{gate} can be scaled from 6 nm to 4.05 nm, resulting in an ultimately small FP of 17 nm, which is close to the projected ultimately small GP of 18 nm discussed in the previous section, for L_{sp} = 6 nm. Thus it is very interesting and important to point out that, at small enough scales, FP actually affects DSDT and vice versa, since L_{sp} constrains the minimum t_{gate} and therefore FP.

As Fig. 3.7 shows, I_{ON} is higher with EOT = 1 nm and ε_{gd} = 3.9 (SiO_2) than with EOT = 0.5 nm and ε_{gd} = 23 (HfO_2) (298 μA/μm vs. 236 uA/μm, without silicide gating). On the other

hand, t_{gd} is smaller (1 nm vs. 2.95 nm), meaning that t_{gate} is larger for constant FP with SiO_2 as the gate dielectric. As a result, the superior gate dielectric solution will be determined not by I_{ON} but instead by the trade-off between maximizing the electrostatic effect of the gate sidewall on improving *SS* and I_{ON} (i.e., smaller t_{gd}) and minimizing the gate sidewall area and therefore C_{GG} (i.e., larger t_{gd}). Thus it would seem that the only case where high-k gate dielectrics may offer a performance advantage, or even equivalent performance, is not only with lower EOT but also very small FP, where the reduction in I_{ON} is offset by the reduction in gate sidewall area and therefore C_{GG}. (This C_{GG} reduction would also have to offset the C_{GG} increase from the lower EOT). This is discussed in detail in Section 3.8.

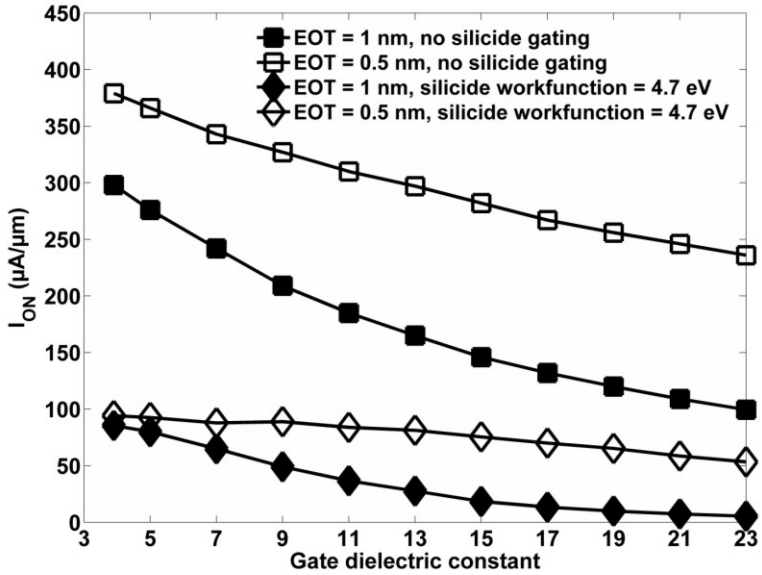

Fig. 3.7. I_{ON}vs. ε_{gd} with and without silicide gating, for $L_{sp} = L_{SDE} = t_{gate}$ = 6 nm, SBH = 0 eV, and HfO_2 spacers with ε_{spacer} = 23.

3.7 Effect of Silicide Gating

In an integrated circuit, double-gate FinFET devices would be electrically connected to other devices, either by strapping multiple source/drain fins together or by accessing the source/drain contact regions with contact vias. Consequently, device design optimization for a fully-integrated structure requires consideration of the silicide gating effect [8] which exists when the work function of the "bulk" silicide (~ 4.7 eV for NiSi), ϕ_M, differs from that of the SDE and channel regions (irrespective of the effective work function at the contact, which may be modified due to dopant segregation). This results in capacitive coupling between the flared silicide source/drain region and the channel and SDE regions through the gate-sidewall spacer and gate dielectric. As a result, there are serious implications for MOSFET design in the DSDT regime, as Fig. 3.7 shows, because both the gate and the silicide compete to control the potential in and near the channel. To model this effect, the structure in Fig. 3.1 is modified by adding blocks of metal adjacent to the sidewall spacers at the source and drain regions, running the full height of the structure, and ϕ_M is varied independently from the contact work function (which is still treated with SBH = 0 eV).

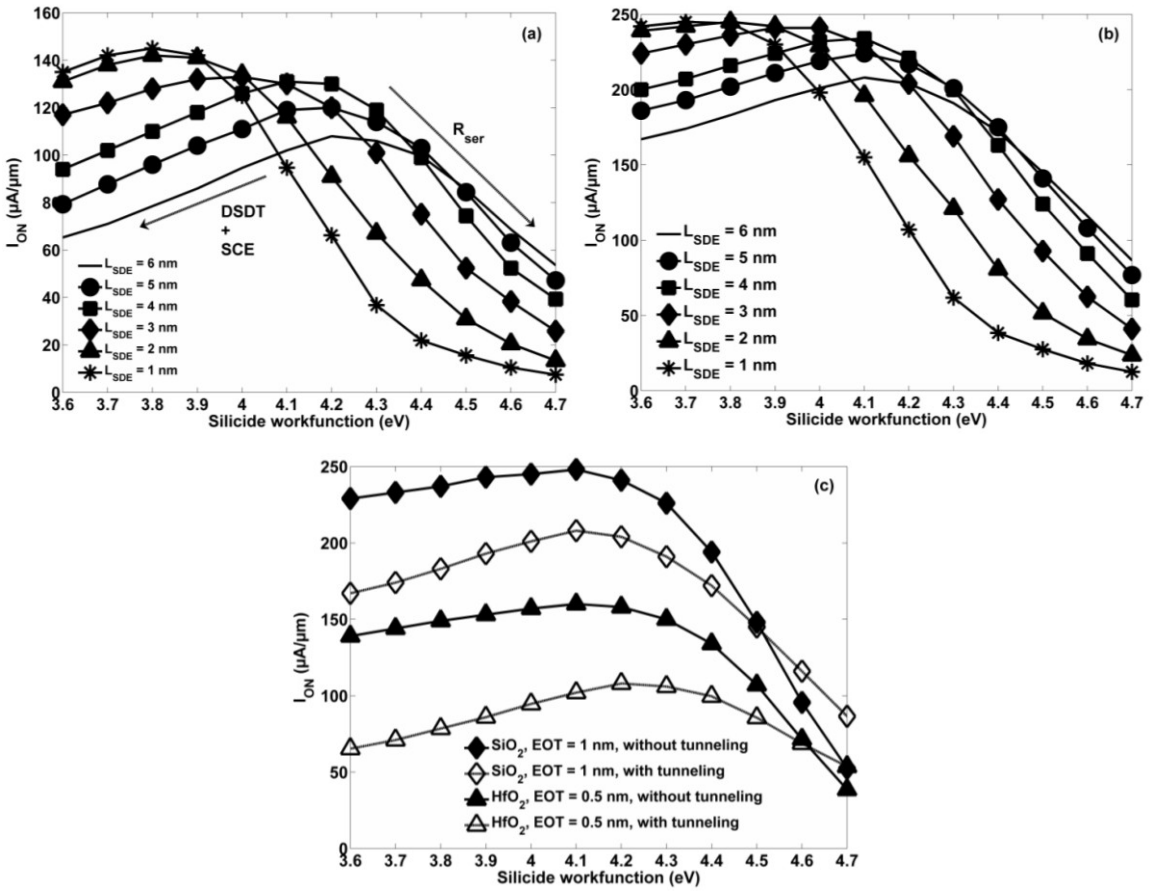

Fig. 3.8. I_{ON} vs. ϕ_M for varying L_{SDE}, with $L_{sp} = t_{gate}$ = 6 nm, L_{SD} = 0, SBH = 0 eV, HfO_2 spacers with ε_{spacer} = 23, and (a) EOT = 0.5 nm (HfO_2 with ε_{gd} = 23), (b) EOT = 1 nm (SiO_2), (c) both gate dielectic cases with and without tunneling and L_{SDE} = 6 nm.

As Fig. 3.8 shows, there exists an optimal ϕ_M, which depends on L_{SDE}. (Here L_{SD} = 0, meaning the peak of the Gaussian SDE profile remains at the SB contact and $L_{SDE} < L_{sp}$ results in a gate underlapped structure.) A high ϕ_M helps to shut the device off by increasing L_{elec}, but limits I_{ON} due to SDE depletion, while a low ϕ_M works in the opposite way. This is much like the SBH trade-off mentioned in Section 3.4, except that, here, SDE depletion is induced by fringing fields extending from the flared silicide. As L_{SDE} drops, L_{elec} increases, requiring less assistance from silicide gating to deplete the underlapped region in the off state. Also, due to the underlapped nature of such a structure where the average SDE doping under the spacer is lower, high ϕ_M values primarily increase parasitic resistance, and so the optimal ϕ_M drops with L_{SDE}. However, as Fig. 3.8(c) shows, the optimal I_{ON} for a given L_{SDE} remains significantly limited by DSDT. It is also important to note that the optimal I_{ON} in Figs. 3.8(a) and (b) is still lower than if silicide gating were not modeled (145 μA/μm vs. 236 μA/μm for HfO_2 with EOT = 0.5 nm and 245 μA/μm vs. 298 μA/μm for SiO_2 with EOT = 1 nm). This means that silicide gating is still significant, even in the optimized case. Although, for SiO_2 gate dielectric with higher EOT, the optimal I_{ON} comes much closer to the "ideal" I_{ON} where silicide gating is ignored (82.1% vs. 61.5%), suggesting that that t_{gd} matters much more than EOT in the DSDT regime, as the lower portion of the gate sidewall must efficiently couple to the SDE regions to maximize gate control.

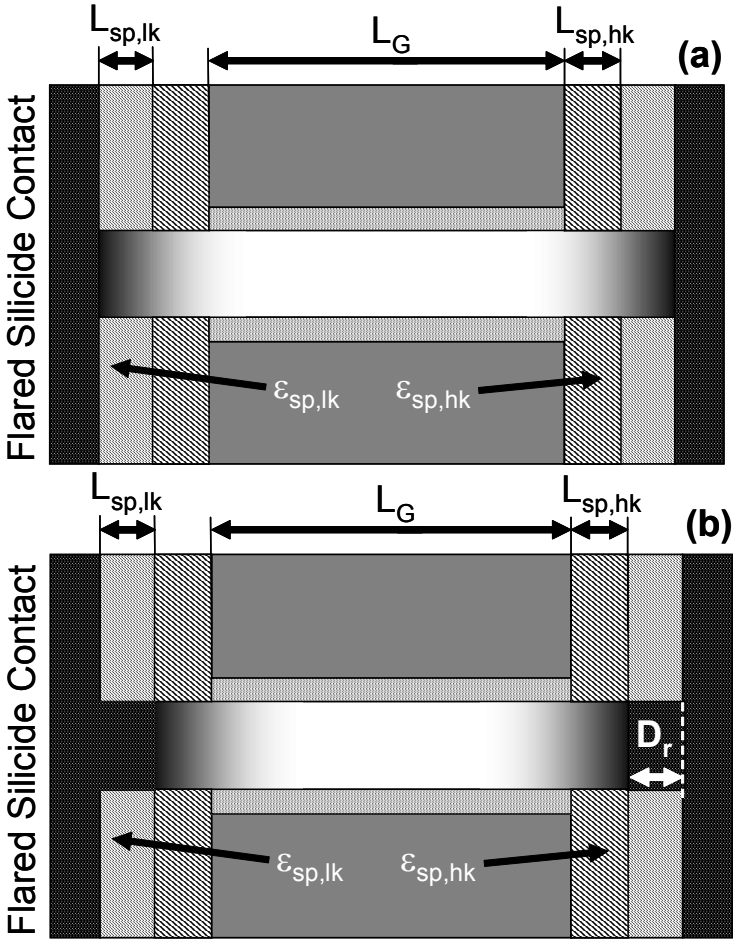

Fig. 3.9. Schematic cross-section of the (a) dual spacer (DS) DSS MOSFET and (b) recessed strap (RS) DSS MOSFET with $D_r = L_{sp,lk}$. All parameters listed in Fig. 3.1 also apply here, except that $L_{SD} = 0$.

An alternative to dual work function silicides for CMOS, which Fig. 3.8 implies, is to fit two spacers into the same space as the single spacer discussed to this point, whereby the outer spacer has a lower dielectric constant. This dual spacer (DS) DSS structure is shown in Fig. 3.9(a) and is very similar to the recessed strap (RS) DSS structure discussed in Chapter 2 and shown in Fig. 3.9(b), but there is a minor difference in that the silicide region does not protrude under the spacers. Although the RS DSS structure was shown in Chapter 2 to effectively eliminate silicide gating for a sufficiently large recess depth D_r, as the flared silicide region is physically pulled away from the contact region, the source-to-drain silicide spacing is closer for the same total spacer length $L_{sp,total}$ (combined low-k and high-k spacer lengths, or $L_{sp,lk}$ and $L_{sp,hk}$, respectively). This results in a lower maximum L_{elec}, which is problematic in the DSDT regime. This is essentially the same exact thing as setting L_{SD} to $L_{sp,lk}$ instead of zero for the DS structure in Fig. 3.9(a), except that now the uniformly doped portion of the SDE region is replaced with metal silicide. The implication here is that L_{SDE} must be scaled with $L_{sp,hk}$ in order to keep $L_{eff} \geq L_G$ for the RS DSS structure at constant $L_{sp,total}$ (which is not necessary for the DS DSS structure). This results in a smaller $L_{sp,lk}$ design space and a lower maximum achievable I_{ON}, as indicated in Figs. 3.2 (a) and (b) for decreasing L_{SDE} at constant L_{eff} and shown explicitly in Figs. 3.10 (a) and (b).

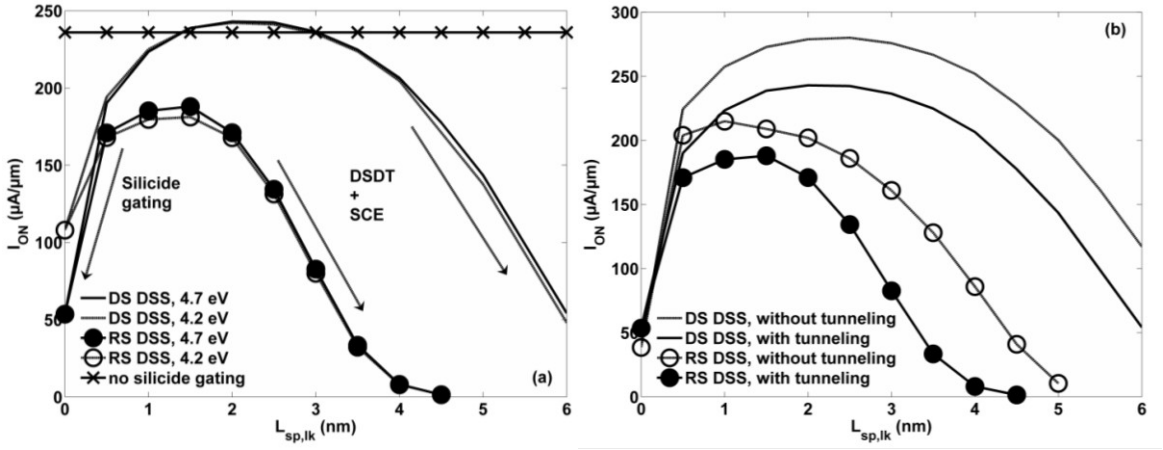

Fig. 3.10. I_{ON} vs. $L_{sp,lk}$ for the DS DSS and RS DSS MOSFETs shown in Fig. 3.12, with (a) varying ϕ_M and (b) ϕ_M = 4.7 eV with and without tunneling. $L_{sp,total} = t_{gate}$ = 6 nm, $\varepsilon_{sp,hk}$ = 23, $\varepsilon_{sp,lk}$ = 2, ε_{gd} = 23 with EOT = 0.5 nm, and SBH = 0 eV. For the DS structure, L_{SDE} = 6 nm, while for the RS structure, $L_{SDE} = L_{sp,hk}$.

Since for the same $L_{sp,total}$ the high-k spacer must be smaller, in order to accommodate the low-k outer spacer (which is treated here as an inter-layer dielectric or ILD, with $\varepsilon_{sp,lk}$ = 2 according to the ITRS projections in this L_G regime [9]), the optimization path for a given EOT and ε_{gd} involves balancing the decrease in L_{elec} and the increase in R_{ser} by L_{SDE} tuning (with L_{SD} = 0), while also varying the lengths and dielectric constants of the high-k ($\varepsilon_{sp,hk}$) and low-k ($\varepsilon_{sp,lk}$) spacers to minimize silicide gating. For a given ϕ_M and $L_{sp,total}$, the DS trade-off is that $L_{sp,hk}$ must be as large as possible to maximize L_{elec} and therefore gate control; however, this also increases the flared silicide coupling (or via coupling for planar structures) to the SDE and body regions, reducing I_{ON}. Increasing $L_{sp,lk}$ reduces this coupling, but at the cost of reduced L_{elec}, increased DSDT and SCE, and reduced I_{ON}. Fig. 3.10 shows an example of this trade-off, for HfO_2 (EOT = 0.5 nm) gate dielectric and $L_{SDE} = L_{sp,total}$ = 6 nm. In this case, $L_{sp,lk}$ does not have to be very large (< 1 nm) to effectively eliminate silicide gating in both the DS and RS DSS structures, since beyond this point I_{ON} shows little if any dependence on ϕ_M. In fact, the DS approach is so effective at reducing silicide gating that the optimal I_{ON} at $L_{sp,lk}$ = 2 nm not only is improved over the single high-k spacer with optimized ϕ_M ($L_{sp,lk}$ = 0 and ϕ_M = 4.2 eV), but also is higher than what is modeled when silicide gating is ignored entirely, as Fig. 3.10(a) shows. (Although not shown here, this was also observed for SiO_2 with EOT = 1 nm.)

Thus, by decoupling the flared silicide region with a low-k outer spacer, device performance becomes independent of ϕ_M and, by definition, variations in ϕ_M due to random variations in silicide grain orientation [23]. Similar to the single spacer case (Fig. 3.8), though, the optimal I_{ON} remains DSDT-limited in both the DS and RS cases, as Fig. 3.13(b) shows. Further reduction in the influence of DSDT for the same $L_{sp,total}$ would require re-optimizing L_{SDE}, since now two spacers are used instead of one.

Although not shown here, the optimal $L_{sp,lk}$ was found to be ~ (1/3)*$L_{sp,total}$, mostly independent of L_{SDE}. The optimal L_{SDE} ~ 0.5-0.6*$L_{sp,total}$ (Fig. 3.11). Although shrinking L_{SDE} allows for a larger $L_{sp,lk}$ design space due to the increase in L_{elec}, there is a R_{ser} trade-off due to the lower average doping under the spacer, meaning that gate fringing fields through the high-k spacer are required to offset the R_{ser} increase. As a result, the optimal $L_{sp,lk}$ for a given $L_{sp,total}$

does not change by much as L_{SDE} is scaled. Likewise, the optimal $L_{sp,total}$ does not change when L_{SDE} scaling is considered, since the reduced source-to-drain contact spacing reduces L_{elec} by more than the increase offered with L_{SDE} scaling, as Fig. 3.11 shows. Regardless, it is clear that a combined optimization of sidewall spacer and SDE design can substantially reduce the influence of DSDT on I_{ON}, whereby DSDT now only just begins to show up at $L_{sp,total}$ = 4-4.5 nm (rather than 6 nm from before) and does not result in a large I_{ON} degradation for smaller $L_{sp,total}$. From the GP and FP discussions earlier, it would seem that optimized device design would allow both to be reduced from 18 nm and 17 nm, respectively, to ~ 14 nm and 13 nm, respectively, before DSDT influences device performance. SCE reduces I_{ON} as $L_{sp,total}$ drops from 6 nm, however, meaning that ultimate circuit scaling in optimized devices may well be limited by SCE and not DSDT, even for extremely small L_G. This conclusion arises specifically from the need to investigate and minimize the effect of silicide gating at such small scales.

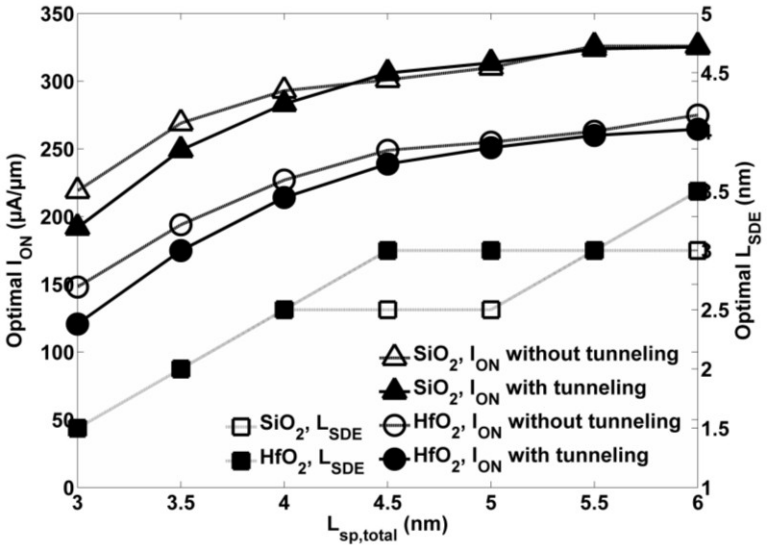

Fig. 3.11. DS DSS Optimal I_{ON} with and without tunneling vs. $L_{sp,total}$ with corresponding L_{SDE} for HfO_2 (ε_{gd} = 23, EOT = 0.5 nm) and SiO_2 (EOT = 1 nm) gate dielectrics. $\varepsilon_{sp,hk}$ = 23, $\varepsilon_{sp,lk}$ = 2, $L_{sp,lk}$ = (1/3)*$L_{sp,total}$, ϕ_M = 4.7 eV, t_{gate} = $L_{sp,total}$, and SBH = 0 eV.

3.8 Delay Optimization

Although it has been shown thus far that an effective switch can be made, at least in theory, at L_G = 3 nm and with a reasonable assumption for t_{body}, it is not yet clear whether, for the corresponding GP or total device length L_{total} = L_G + 2*$L_{sp,total}$, this structure has the lowest intrinsic delay. Although the purpose of the high-k spacer is to increase L_{elec} and therefore I_{ON}, this can also be achieved by displacing the high-k spacer with gate electrode material (*i.e.*, increasing L_G at the cost of $L_{sp,hk}$ while keeping $L_{sp,total}$ constant, meaning the maximum L_G = 3 nm + 2*$L_{sp,hk}$). The trade-off here is increased fringing capacitance between the gate sidewall and the flared silicide region, but it may be that optimal circuit delay in the DSDT regime constrains L_G to larger or perhaps smaller values than what was modeled up to this point. To investigate this, the intrinsic delay C_{GG}*V_{DD}/I_{ON} is modeled for the DS DSS structure in Fig. 3.9(a). Fig. 3.12 shows C_{GG}*V_{DD}/I_{ON} vs. L_G for constant L_{total}, where $L_{sp,lk}$ and L_{SDE} are optimized using L_G = 3 nm as in Section 3.7. The quantity t_{gate} + t_{gd} is also held constant at each

L_{total} and scaled with L_{total} (*i.e.,* $t_{gate} + t_{gd} = 0.5*[L_{total} - 3\text{ nm}] + 1\text{ nm}$), meaning that both FP and GP are scaled together in Fig. 3.12, as per the FP scaling discussion in Section 3.6. Thus for a given L_{total} (which leads to GP), FP is at its lowest value for keeping DSDT and SCE at bay, and is expressed as $L_{total} + t_{body} - 1$ nm.

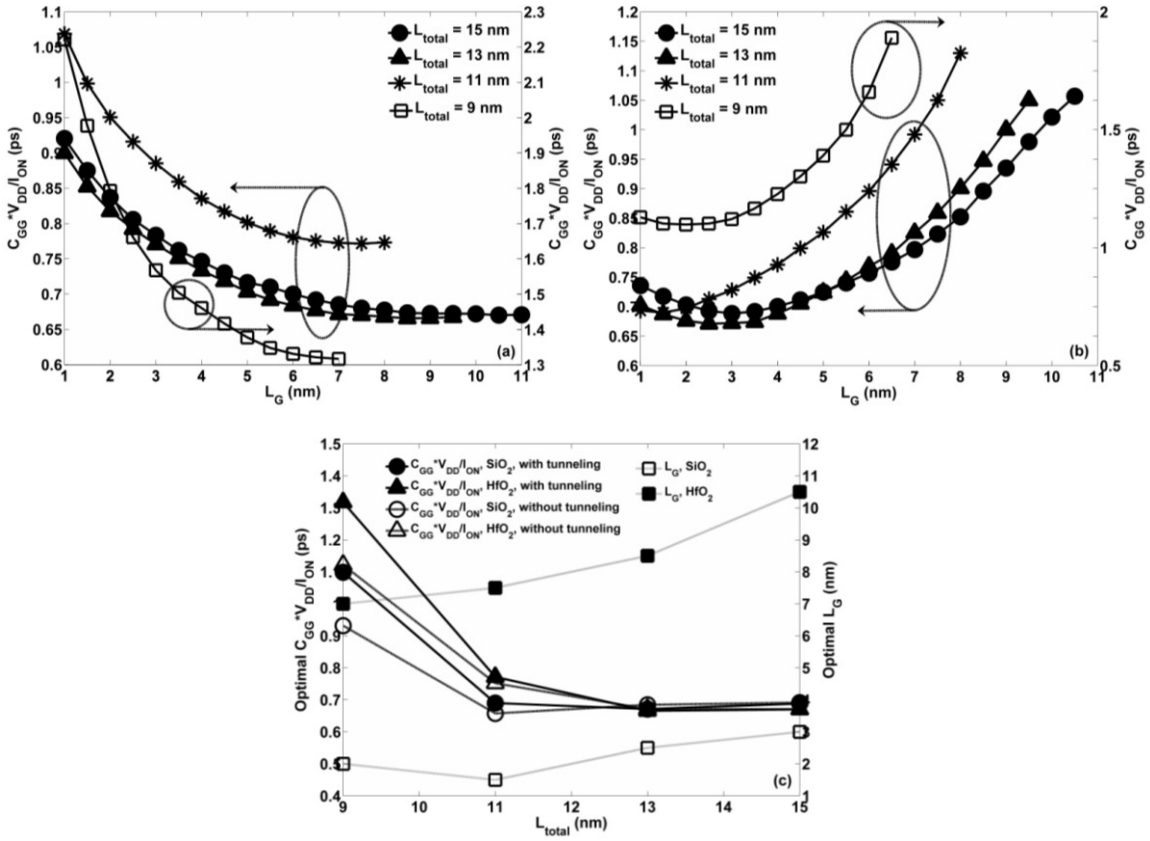

Fig. 3.12. (a) Intrinsic delay vs. L_G for HfO_2 gate dielectric with EOT = 0.5 nm, (b) intrinsic delay vs L_G for SiO_2 gate dielectric with EOT = 1 nm, and (c) optimal intrinsic delay vs. L_{total} with and without tunneling for both gate dielectric cases, along with the corresponding optimal L_G. In all cases, the L_{SDE}- and $L_{sp,lk}$-optimized DS DSS structure is used with $\phi_M = 4.7$ eV, SBH = 0 eV, $\varepsilon_{sp,hk} = 23$, and $\varepsilon_{sp,lk} = 2$.

For HfO_2 (EOT = 0.5 nm), delay drops as L_G increases, due to the increase in L_{elec} and reduction in SCE, which apparently more than offsets the increase in C_{GG} as the gate-to-flared silicide spacing is reduced. (Here the optimal I_{ON} is achieved with $L_G = 3\text{ nm} + 2*L_{sp,hk}$, also expressed as $1\text{ nm} + (2/3)*L_{total}$, which means no high-k spacer, although the optimal delay requires L_G slightly smaller than this to reduce the gate-to-flared-silicide coupling.) For SiO_2 (EOT = 1 nm), the exact opposite trend is observed and so the situation is more complicated. In this case, the physical separation between the lower portion of the gate sidewall and the SDE region is also 1 nm (and increases moving up along the gate sidewall), but along the gate sidewall the capacitive coupling takes place through the high-k spacer. As a result, the "sidewall EOT" along the lower portion of the gate sidewall is much less than 1 nm, so that the gate sidewall-to-SDE coupling plays a stronger role in current modulation than the gate electrode-to-channel coupling. Thus, as L_G shrinks, the gate sidewall contribution to total gate control increases, which increases C_{GG}, but also I_{ON}, for a net reduction in delay. Only for very small L_G does I_{ON} drop and delay increase. This is due in part to R_{ser}, as Fig. 3.12(c) shows the optimal L_G

dropping as L_{total} drops from 15 nm to 11 nm (*i.e.*, as L_G shrinks for constant L_{total}, the amount of underlapped SDE region increases and so R_{ser} increases), and in part to SCE, whereby the optimal L_G increases again as L_{total} drops from 11 nm to 9 nm.

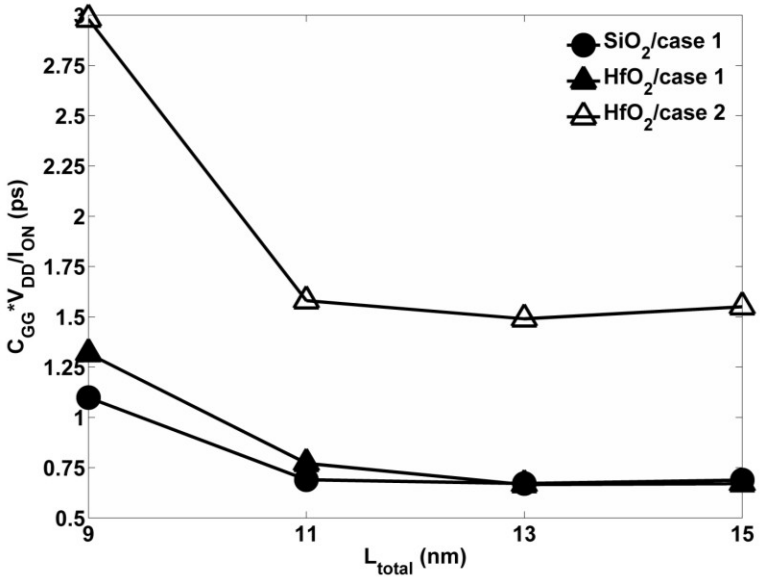

Fig. 3.13. Intrinsic delay vs. L_{total}, comparing the optimized cases in Fig. 3.12(c), or case 1, and case 2, where FP is scaled less aggressively such that t_{gate} matches that of SiO_2/case1.

It is clear from Fig. 3.12 that similar delay can be achieved using very different solutions for L_G and the gate dielectric, and that device performance depends much more on the source-to-drain contact spacing (L_{total}) than it does on L_G. For optimized devices, only for $L_{total} \leq 11$ nm does DSDT significantly affect delay, and even in this case its contribution is much smaller than that of SCE. Thus it would seem fair to consider t_{body} scaling at this point. However, scaling t_{body} below 3 nm (not shown) does not improve much on what Fig. 3.12(c) already shows for L_{total} = 9 nm. For SiO_2 with EOT = 1 nm (which scales better than HfO_2 with EOT = 0.5 nm, according to Fig. 3.12), intrinsic delay drops from 1.1 ps to 0.97 ps as t_{body} is reduced from 3 nm to 2.5 nm, and then increases as t_{body} is reduced below 2.5 nm. This is because mobility degradation due to surface scattering at the gate dielectric interfaces offsets any further gain achieved in gate control.

It would therefore seem that the device designer is left with two options for minimizing delay for small GP and therefore L_{total}. One option is to use SiO_2 gate dielectric with relaxed EOT (but thinner t_{gd}). (Here, C_{GG} is higher but low delay is achieved through an I_{ON} advantage.) However, the challenge here is patterning much smaller L_G, perhaps with spacer lithography [24]. The second option is to use larger L_G (easier to pattern), but contend with the challenge of manufacturing high-k gate dielectrics with extremely low EOT. (Here, I_{ON} is lower but low delay is achieved through a C_{GG} advantage.) Another problem here is that FP will be difficult to scale, since t_{gd} is an appreciable portion of FP, even for HfO_2 with EOT = 0.5 nm. For example, at L_{total} = 9 nm, $t_{gate} + t_{gd} = L_{sp0} + 1$ nm = 4 nm, where L_{sp0} is $L_{sp,total}$ when L_G = 3 nm (recalling Fig. 3.5). In this case, t_{gd} = 2.95 nm, meaning that the remaining gap between the fins that is to be filled with gate electrode material is $2*t_{gate}$ = 2.1 nm. From a process standpoint, this may be a difficult gap to fill with metal, notwithstanding the challenges imposed in effectively accessing this region for gate work function engineering. One could relax FP to alleviate this problem, but

this limits circuit density scaling and increases C_{GG} (due to the added gate sidewall area) and therefore delay significantly, since I_{ON} saturates for large enough t_{gate} and therefore FP (Fig. 3.5). This is shown in Fig. 3.13, where "case 1" represents the conditions in Fig. 3.12(c) and "case 2" is where FP is scaled less aggressively such that t_{gate} is the same as it is for SiO_2/case 1.

One could make the argument that, although SiO_2/case 1 is more scalable than HfO_2/case 1, its use of a larger high-k inner spacer would result in higher Miller capacitance C_M and therefore a larger Miller effect during inverter switching. To address this question, mixed-mode inverter simulations were performed for three cases: (a) SiO_2/case 1; (b) HfO_2/case 1; (c) SiO_2/case 1 with a single high-k spacer, each at L_{total} = 11 nm and 15 nm. Mixed mode circuit simulations do not permit the use of the 1-D Schrödinger quantization or tunneling models, so the simulation was performed using the DGM quantization model and without any tunneling current. The exclusion of tunneling in the simulation is reasonable, considering the minimal DSDT in the optimized devices shown previously. The PMOS device was simulated using the exact same geometries and doping concentrations as the NMOS device. For case (c) (SiO_2/case 1 with a single high-k spacer), silicide gating is significant since there is now low-k outer spacer to decouple the flared silicide region from the SDE. As such, the NMOS and PMOS silicide workfunctions are set to the conduction and valence band edges, respectively, to maximize I_{ON} in light of the silicide gating effect. The PMOS width was increased appropriately to result in matched NMOS and PMOS I_{ON} and the load capacitance at the inverter output was set to 1 fF/μm. The simulation results are shown in Fig. 3.14, with the extracted rise and fall propagation delays (τ_{pr} and τ_{pf}, respectively) shown in Table 3.1.

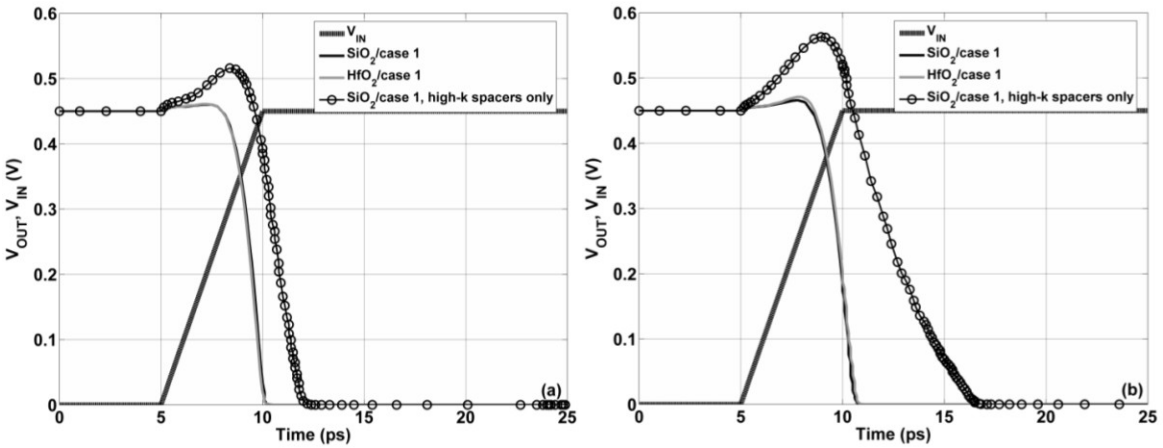

Fig. 3.14. Mixed mode inverter simulation results for various gate stack designs and (a) L_{total} = 15 nm and (b) L_{total} = 11 nm. Only the pull-down operation is shown here, since the pull-up operation is symmetric.

As Fig. 3.14 shows, the Miller effect for SiO_2/case 1 is the same as or lower than that of HfO_2/case 1. For the split with high-k spacers only (i.e., case (c)), the Miller effect is very large, due to increased fringing capacitance between the gate sidewall and flared silicide regions. For cases (a) and (b), the propagation delays are very similar, as Table 3.1 shows. Thus it would seem that high-k spacers, when properly implemented with low-k outer spacers to minimize C_M, do not impose any performance penalty. As such, it is impossible at this point to predict with certainty which gate dielectric solution will prove superior near the end of the CMOS technology roadmap, since the answer relies upon the developmental limits of several process technologies such as gate patterning and high-k dielectric etching. Despite this ambiguity, it is worth noting

that high-k gate dielectrics are not a ubiquitous solution, especially in the DSDT regime where fringing field effects play such an important role, and that high-k dielectrics in general may be more useful as something other than the gate dielectric (this theme is further explored in Chapter 5).

Table 3.1. Extracted propagation delays from mixed mode inverter simulations

	L_{total} = 15 nm		L_{total} = 11 nm	
	τ_{pr} (ps)	τ_{pf} (ps)	τ_{pr} (ps)	τ_{pf} (ps)
SiO_2/case 1	1.78	1.77	1.96	2.20
HfO_2/case 1	1.59	1.55	2.07	2.21
SiO_2/case 1, high-k spacers only	2.53	2.95	4.82	4.87

3.9 Summary

DSDT was studied through TCAD simulation as a potential limiting factor to DG MOSFET scalability in the sub-10 nm regime. It was shown that fringing field effects from the flared silicide and gate sidewall regions play a critical role in device optimization in this regime, and that DSDT can be suppressed so as to have zero or near-zero influence on device performance for $L_{total} \geq 11$ nm. As L_{total} is reduced below 11 nm, DSDT becomes significant, but conventional thermal leakage remains much greater, thus suggesting that DSDT will not be a limiting factor to CMOS scaling in optimized devices. Perhaps the most profound finding here is that high-k gate dielectrics have utility as something other than the gate dielectric, and that the gate dielectric thickness is more important than EOT in the sub-10 nm regime. As it turns out, conventional SiO_2 gate dielectrics exhibit superior scalability over HfO_2 gate dielectrics, so long as a dual high-k/low-k sidewall spacer is employed. This dual spacer design comes at no cost to the Miller effect in the transient behavior of these devices, since the low-k outer spacer decouples the gate sidewall/high-k inner spacer from the flared silicide region.

3.10 References

[1] H. Kawamura, T. Sakamoto, T. Baba, "Observation of source-to-drain direct tunneling current in 8 nm gate electrically variable shallow junction metal-oxide-semiconductor field-effect transistors," *Appl. Phys. Lett.*, vol. 76, no. 25, pp. 3810-3812, Jun. 2000.

[2] J. R. Watling, A. Asenov, A. R. Brown, A. Svizhenko, M. P. Anantram, "Direct Source-to-Drain Tunneling and its Impact on the Intrinsic Parameter Fluctuations in nanometer scale Double Gate MOSFETs," *Nanotech*, vol. 2, pp. 202-205, 2003.

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

[4] K. D. Cantley, Y. Liu, H. S. Pal, T. Low, S. S. Ahmed, M. S. Lundstrom, "Performance Analysis of III-V Materials in a Double-Gate nano-MOSFET," *IEDM Tech. Dig.*, 2007, pp. 113-116.

[5] Q. Rafhay, R. Clerc, G. Ghibaudo, G. Pananakakis, "Impact of source-to-drain tunneling on the scalability of arbitrary oriented alternative channel material nMOSFETs," *Solid-State Electronics*, vol. 52, pp. 1474-1481, 2008.

[6] *User's Manual for Sentaurus Device*, Synopsys Co., Mountainview, CA.

[7] R. A. Vega, T.-J. King Liu, "A Comparative Study of Dopant-Segregated Schottky and Raised Source/Drain Double-Gate MOSFETs," *IEEE Trans. Elec. Dev.*, vol. 55, no. 10, pp. 2665-2677, Oct. 2008.

[8] R. A. Vega, T.-J. King Liu, "Three-Dimensional FinFET Source/Drain and Contact Design Optimization Study," *IEEE Trans. Elec. Dev.*, (to be published, July 2009).

[9] International Technology Roadmap for Semiconductors (ITRS). [Online]. Available: http://public.itrs.net

[10] R. Venugopal, Z. Ren, M. S. Lundstrom, "Simulating Quantum Transport in Nanoscale MOSFETs: Ballistic Hole Transport, Subband Engineering and Boundary Conditions," *IEEE Trans. Nanotech.*, vol. 2, no. 3, pp. 125-143, Sept. 2003.

[11] H. Lee, L.-E. Yu, S.-W. Ruy, J.-W. Han, K. Jeon, D.-Y. Jang, K.-H. Kim, J. Lee, J.-H. Kim, S. C. Jeon, G. S. Lee, J. S. Oh, Y. C. Park, W. H. Bae, H. M. Lee, J. M. Yang, J. J. Yoo, S. I. Kim, Y.-K. Choi, "Sub-5nm All-Around Gate FinFET for Ultimate Scaling," *VLSI Tech. Dig.*, 2006, pp. 58-59.

[12] N. Mise, S. Migita, Y. Watanabe, H. Satake, T. Nabatame, A. Toriumi, "(111)-Faceted Metal Source and Drain for Aggressively Scaled Metal/High-k MISFETs," *IEEE Trans. Elec. Dev.*, vol. 55, no. 5, pp. 1244-1249, May 2008.

[13] T.-Y. Liow, K.-M. Tan, R. T. P. Lee, M. Zhu, B. L.-H. Tan, G. S. Samudra, N. Balasubramanian, Y.-C. Yeo, "5 nm Gate Length Nanowire-FETs and Planar UTB_FETs with Pure Germanium Source/Drain Stressors and Laser-Free Melt-Enhanced Dopant (MeltED) Diffusion and Activation Technique," *VLSI Tech. Dig.*, 2008, pp. 36-37.

[14] K. Shenai, E. Sangiorgi, R. M. Swanson, K. C. Saraswat, R. W. Dutton, "Modeling and Characterization of Dopant Redistributions in Metal and Silicide Contacts," *IEEE Trans. Elec. Dev.*, vol. 32, no. 4, pp. 793-799, Apr. 1985.

[15] M. Ono, A. Nishiyama, M. Koyama, "Degradation of current drivability of Schottky barrier source/drain transistors induced by high-k gate dielectrics and possible measures to suppress the phenomenon," *Solid State Electronics*, vol. 50, pp. 788-794, 2006.

[16] T. Miyashita, K. Ikeda, Y. S. Kim, T. Yamamoto, Y. Sambonsugi, H. Ochimizu, T. Sakoda, M. Okuno, H. Minakata, H. Ohta, Y. Hayami, K. Ookoshi, Y. Shimamune, M. Fukuda, A. Hatada, K. Okabe, T. Kubo, M. Tajima, T. Yamamoto, E. Motoh, T. Owada, M. Nakamura, H. Kudo, T. Sawada, J. Nagayama, A. Satoh, T. Mori, A. Hasegawa, H. Kurata, K. Sukegawa, A. Tsukune, S. Yamaguchi, K. Ikeda, M. Kase, T. Futatsugi, S. Satoh, T. Sugii, "High-Performance and Low-Power Bulk Logic Platform Utilizing FET Specific Multiple-Stressors with Highly Enhanced Strain and Full-Porous Low-k Interconnects for 45-nm CMOS Technology," *VLSI Tech. Dig.*, pp. 251-254, 2007.

[17] A. B. Sachid, R. Francis, M. S. Baghini. D. K. Sharma, K.-H. Bach, R. Mahnkopf, V. R. Rao, "Sub-20 nm Gate Length FET Design: Can High-k Spacers Make a Difference?," *IEDM Tech. Dig.*, pp. 697-700, 2008.

[18] R. A. Vega, "Schottky field effect transistors and Schottky CMOS circuitry," M.S. thesis, Dept. Microelectron. Eng., Rochester Inst. Technol., Rochester, NY, 2006.

[19] X. Chen, S. Samavedam, V. Narayanan, K. Stein, C. Hobbs, C. Baiocco, W. Li, D. Jaeger, M. Zaleski, H. S. Yang, N. Kim, Y. Lee, D. Zhang, L. Kang, J. Chen, H. Zhuang, A. Sheikh, J. Wallner, M. Aquilino, J. Han, Z. Jin, J. Li, G. Massey, S. Kalpat, R. Jha, N. Moumen, R. Mo, S. Kirshnan, X. Wang, M. Chudzik, M. Chowdhury, D. Nair, C. Reddy, Y. W. Teh, C. Kothandaraman, D. Coolbaugh, S. Pandey, D. Tekleab, A. Thean, M. Sherony, C. Lage, J. Sudijono, R. Lindsay, J. H. Ku, M. Khare, A. Steegen, "A Cost Effective 32nm High-K/ Metal Gate CMOS Technology for Low Power Applications with Single-Metal/Gate-First Process," *VLSI Tech. Dig.*, 2008, pp. 88-89.

[20] S. Kubicek, T. Schram, E. Rohr, V. Paraschiv, R. Vos, M. Demand, C. Adelmann, T. Witters, L. Nyns, A. Delabie, L.-A. Ragnarsson, T. Chiarella, C. Kerner, A. Mercha, B. Parvais, M. Aoulaiche, C. Ortolland, H. Yu, A. Veloso, L. Witters, R. Singanamalla, T. Kauerauf, S. Brus, C. Vrancken, V. S. Chang, S.-Z. Chang, R. Misuhashi, Y. Okuno, A. Akheyar, H.-J. Cho, J. Hooker, B. J. O'Sullivan, S. Van Elshocht, K. De Meyer, M. Jurczak, P. Absil, S. Biesemans, T. Hoffman, "Strain enhanced Low-V_t CMOS featuring La/Al-doped HfSiO/TaC and 10ps Invertor Delay," *VLSI Tech. Dig.*, 2008, pp. 130-131.

[21] C. Auth, A. Cappellani, J.-S. Chun, A. Dalis, A. Davis, T. Ghani, G. Glass, T. Glassman, M. Harper, M. Hattendorf, P. Hentges, S. Jaloviar, S. Joshi, J. Klaus, K. Kuhn, D. Lavric, M. Lu, H. Mariappan, K. Mistry, B. Norris, N. Rahhal-orabi, P. Ranade, J. Sandford, L. Shifren, V. Souw, K. Tone, F. Tambwe, A. Thompson, D. Towner, T. Troeger, P. Vandervoorn, C. Wallace, J. Wiedemer, C. Wiegand, "45nm High-k + Metal Gate Strain-Enhanced Transistors," *VLSI Tech. Dig.*, 2008, pp. 128-129.

[22] Q. Chen, L. Wang, J. D. Meindl, "Fringe-induced barrier lowering (FIBL) included threshold voltage model for double-gate MOSFETs," *Solid-State Electronics*, vol. 49, pp. 271-274, 2005.

[23] H. Dagour, K. Endo, V. De, K. Banerjee, "Modeling and Analysis of Grain-Orientation Effects in Emerging Metal-Gate Devices and Implications for SRAM Reliability," *IEDM Tech. Dig.*, 2008, pp. 705-708.

[24] Y.-K. Choi, T.-J. King, C. Hu, "A Spacer Patterning Technology for Nanoscale CMOS," *IEEE Trans. Elec. Dev.*, vol. 49, no. 3, pp. 436-441, Mar. 2002.

Chapter 4

The Effect of Random Dopant Fluctuations on Specific Contact Resistivity

4.1 Introduction

As the MOSFET is scaled into the nanometer regime, source/drain series resistance and contact resistance become an increasingly large fraction of its total on-state resistance. This is particularly true for a thin-body structure (*e.g.*, ultra-thin-body MOSFET, FinFET, or nanowire MOSFET) that may be necessary to maintain the historical rate of transistor scaling. It has been shown in [1]-[2] that, for a sufficiently high dopant concentration, the specific contact resistivity ρ_c values for NiSi and PtSi contacts to n-type and p-type Si can easily meet the ITRS specifications at the end of the roadmap [3]. However, the effects of random dopant fluctuation (RDF), which become more significant with decreasing contact area (even for doping levels on the order of 10^{20} cm^{-3}), have yet to be quantifed.

For example, a FinFET with 10 nm gate length would have fin(s) approximately 7 nm wide by 21 nm tall in dimension [4]. A Schottky-barrier (SB) depletion region that is 3-nm wide – *e.g.* for a NiSi contact with SB height (SBH) = 0.65 eV, neglecting SB lowering (SBL) -- would nominally contain 44 randomly distributed dopant atoms, for 10^{20} cm^{-3} nominal source/drain dopant concentration. The variation in the number of dopant atoms ($44^{1/2}$) is ~15 % of the nominal value, suggesting that ρ_c would vary by more than 15% due to the exponential dependence of ρ_c on dopant concentration [1]-[2], [5]-[6]. On the other hand, SBL due to heavy doping may reduce the sensitivity of ρ_c to RDF. It is therefore worthwhile to investigate how significant ρ_c variation will become as CMOS technology scaling continues, and to study its dependencies on dopant concentration, SBH, and contact area. In this work, an analytical model is developed to allow for quick and reasonably accurate prediction of RDF effects on ρ_c. This model is calibrated to TCAD simulations using Sentaurus Device [7] and also to previously published data [1]-[2].

4.2 Modeling Approach

The device structure modeled using Sentaurus Device is a simple 2-dimensional (2-D) diode. The length between the contacts is set to 20 nm and the dopant concentration N is set to some nominal value (*e.g.* $1x10^{20}$ cm^{-3}). For each combination of SBH and N values, ρ_c is extracted by

taking the difference in resistance between having an ohmic contact at each end of the device and having one Schottky contact (with SBH varied from 0-0.6 eV) at one end of the device. A continuum doping profile is used for these device simulations, since the TCAD software is not yet capable of accurate electrical simulation with atomistic doping profiles.

Although the results are not shown here, attempts have been made to simulate the effect of RDF on ρ_c for a 3-D diode structure, using the Kinetic Monte Carlo (KMC) simulator in Sentaurus Device. After the device geometry is defined and N is set to a nominal value, the KMC simulator determines a random dopant distribution. The individual dopant atoms are modeled as having nearly-delta-function distributions in continuum mode, with a peak concentration of $> 1\text{x}10^{22}$ cm^{-3}, so that the space between the dopant atoms is treated as undoped. The problem with this approach is that the conductivity of the structure is modeled classically, meaning that the path of least resistance is through the regions that have the highest dopant concentration; however, in reality, the mobile carrier concentration within the structure is uniform while the carrier mobility is non-uniform – being lower near the dopant atoms due to Coulombic effects, and higher in the regions in-between the dopant atoms – so that current flows primarily through the regions that have the lowest dopant concentration. In other words, the I-V simulation does not account for the fact that the structure is in the atomistic regime. As a result, the KMC-simulated resistance is several orders of magnitude higher than that obtained by simulation in continuum mode assuming a uniform doping profile.

This problem can be resolved, in part, by performing a short diffusion simulation (1000 ºC for 1 sec. was found to be sufficient), which spreads out the dopant distributions just enough for the resistance to be similar in both KMC and continuum simulations. However, the free electron concentration remains non-uniform and the simulated ρ_c for very large contact areas (for which RDF is negligible) is larger than the results obtained from the 2-D continuum approach. The structure can be annealed further to form a more uniform dopant (and therefore free electron) distribution, but this approach is somewhat arbitrary and not physically rigorous. Since atomistic doping effects on contact resistance cannot be accurately simulated using the currently available TCAD software, an analytical model for ρ_c is developed in this work and calibrated against 2-D continuum-mode simulations. Then the analytical model is developed further to obtain an expression which quantitatively describes the impact of RDF on ρ_c.

Both the "Schottky" contact model (for thermionic emission over the SB) and a non-local 1-D Schrödinger tunneling model [7] are used, as in [4], [8]-[9]. The electron and hole effective Richardson's constants are set to 112 and 32 $\text{A/cm}^2\text{*K}^2$, respectively. For contact to n-type Si, the electron tunneling mass is set to $0.19*m_0$. For contact to p-type Si, light hole and heavy hole masses of $0.16*m_0$ and $0.49*m_0$ are used, respectively. Since the heavy hole mass is low enough to have a non-negligible effect on ρ_c, the p-type contact simulation is run twice, once for each hole tunneling mass; then the light hole and heavy hole subbands are treated as acting in parallel in order to determine ρ_c.

Quantization effects are excluded in these simulations, since there are multiple ways to achieve the same contact area. For example, a 1-nm-thick SOI layer (which would have significant energy quantization) that is 100 nm wide would have the same contact area as a 10-nm-thick layer that is 10 nm wide. Likewise, the dimensionality of the system and hence its effect on the density of states (DOS) is also neglected here. Using the same example, the former case would have a 2-D DOS, while the latter case would likely have a 1-D DOS. Although DOS does not affect the probability of tunneling through a SB, it does affect the tunneling current because it affects availability of states into which electrons or holes can tunnel. To avoid any ambiguity

due to differing energy quantization and DOS for different structures, this work focuses on basic effects of the contact parameters: contact area, SBH, and dopant concentration. The effects of band-gap narrowing (BGN) due to heavy doping as well as SBL are not explicitly included, since the simulation results are used to calibrate the analytical model for a given SBH and N.

4.3 Analytical Model Derivation

The analytical model for ρ_c variation begins with a modified form of the unified model in [6], which covers the doping range (~ 10^{18} - 10^{21} cm^{-3}) where either tunneling emission (TE) or thermionic field emission (TFE) dominates current flow in the reverse-biased Schottky junction:

$$\rho_c = \rho_0 \exp\left[\frac{q\phi_B}{kT + E_{00}\coth\left(\frac{E_{00}}{kT}\right)}\right] \tag{4.1}$$

where ρ_0 is ρ_c for infinite dopant concentration (1x10^{-9} Ohm-cm^2 in this study), ϕ_B is the contact SBH, k is the Boltzmann constant, T is the absolute temperature of the system (300 K in this study), q is the electronic charge, and E_{00} is defined as

$$E_{00} = \frac{q\hbar}{2}\sqrt{\frac{N}{m^* m_0 \varepsilon}} \tag{4.2}$$

where $\hbar$ is the Reduced Planck's constant, ε is the semiconductor dielectric permittivity, m^* is the effective mass, and m_0 is the electron rest mass. The term E_{00} was originally defined in [10] as a characteristic energy, which has its roots in the WKB approximation. Thus, E_{00} can be expressed in terms of the electric field E:

$$E_{00} = \frac{E\hbar}{2}\sqrt{\frac{q}{2\phi_B m^* m_0}} \tag{4.3}$$

$$E = \sqrt{\frac{2qN\phi_B}{\varepsilon}} \tag{4.4}$$

Equation (4.4) necessarily assumes that the built-in voltage V_{bi} of the SB contact is equal to ϕ_B for sufficiently heavy doping (such that $E_F \sim E_c$). Although this is not precisely the case, it is a reasonable approximation which simplifies the expression for E so that the standard deviation of E, or σ_E, can be easily derived. As for the TCAD simulations used in this study, SBL is not explicitly included in the analytical model. As discussed in [8], modeling of SBL while accounting for image force, interface dipole, and BGN requires an iterative approach. This results in a non-closed-form solution and makes it difficult to develop an analytical model for variations in E (which depend on variations in ϕ_B) and therefore variations in ρ_c. Thus, the

derivation here for σ_E assumes an "average" ϕ_B, or $\phi_{B,avg}$, at the contact interface. This is not to say that SBL must be excluded entirely from the analysis for $\sigma_{\rho c}$, however. The simplest way to account for variations in SBL is to develop a separate solution for SBL, which is then incorporated into the $\sigma_{\rho c}$ model, as will be shown later.

Returning to Equations (4.3) and (4.4), it is very useful to express E_{00} in terms of E, because deriving σ_E (and eventually $\sigma_{\rho c}$) is relatively straightforward and very similar to the derivation shown in [11] for MOSFET threshold voltage fluctuation. The derivation for σ_E is as follows. From Poisson's Equation, E at a Schottky junction is expressed as

$$E = \frac{qNW_D}{\varepsilon} = \frac{qQ}{\varepsilon} \tag{4.5}$$

where Q is the charge density per unit area at the junction. Fluctuation in Q over a small volume of $dxdydz$ is expressed as

$$\Delta Q = \frac{\sqrt{Ndxdydz}}{A} \tag{4.6}$$

where A is the contact area. The effect of ΔQ on the fluctuation in E is

$$\Delta E = \frac{q\sqrt{Ndxdydz}}{A\varepsilon}\left(1 - \frac{x}{W_{D0}}\right) \tag{4.7}$$

where W_{D0} is the depletion width calculated using the nominal value of N and zero-field ϕ_B, or ϕ_{B0}. The variance of E, or σ_E^2, is then obtained by integrating the square of (4.7) over the volume of W_{D0}, as shown in Equation (4.8). The square root of the solution to Equation (4.8) is taken to obtain σ_E, as shown in Equation (4.9).

$$\sigma_E^2 = \left(\frac{q}{A\varepsilon}\right)^2 N \iiint_{volume}\left(1 - \frac{x}{W_{D0}}\right)^2 dxdydz \tag{4.8}$$

$$\sigma_E = \frac{q}{\varepsilon}\sqrt{\frac{NW_{D0}}{3A}} \tag{4.9}$$

$$\sigma_{E00} = \frac{q\hbar}{2\varepsilon}\sqrt{\frac{qNW_{D0}}{6A\phi_B m^* m_0}} \tag{4.10}$$

Replacing E with σ_E (Equation (4.9)) in Equation (4.3) results in σ_{E00}, as Equation (4.10) shows. Now the variation in ρ_c can be modeled using Equation (4.11). Since ρ_c has an exponential dependence on N and ϕ_B, the average ρ_c, or $\rho_{c,avg}$, is calculated as the logarithmic average, as Equation (4.12) shows, where ρ_{nom} is ρ_c calculated without RDF. A limitation of this

model is that, for very small A and low N, $\sigma_{E00} > E_{00}$, meaning that (E_{00} - σ_{E00}) < 0. This will give erroneous results for $\rho_{c,avg}$, which relies in part on (E_{00} - σ_{E00}). This problem is handled with a conditional statement, which replaces (E_{00} - σ_{E00}) with a very low positive number (*e.g.*, 0.0001) when (E_{00} - σ_{E00}) < 0.

$$\rho_c \pm \sigma_{\rho c} = \rho_0 \exp\left[\frac{q\left(\phi_{B,avg} \pm \sigma_{\phi B}\right)}{kT + \left(E_{00} \mp \sigma_{E00}\right)\coth\left(\frac{E_{00} \mp \sigma_{E00}}{kT}\right)}\right] \tag{4.11}$$

$$\rho_{c,avg} = 10^{\log_{10}\left(\rho_{nom} - \sigma_{\rho c}\right) + 0.5 * \log_{10}\left(\frac{\left(\rho_{nom} + \sigma_{\rho c}\right)}{\left(\rho_{nom} - \sigma_{\rho c}\right)}\right)} \tag{4.12}$$

Returning to the issue of accounting for variations in SBL, the easiest approach to achieving a closed-form solution is to fit a polynomial to a self-consistent SBL model that accounts for image force, interface dipole, and BGN. The SBL model is expressed as:

$$\Delta\phi_B = \Delta\phi_{B,image} + \Delta\phi_{B,dipole} = \sqrt{\frac{qE}{4\pi\varepsilon}} + \frac{\beta Q_s \lambda}{\varepsilon}\exp\left[\frac{-x_m}{\lambda}\right] \tag{4.13}$$

where x_m is the position of maximum potential, β is the fraction of dopants in the Si side of the junction that contribute to dipole lowering, and λ is the average Heine tail decay length.

$$x_m = \sqrt{\frac{q}{16\pi\varepsilon E}} \tag{4.14}$$

$$Q_s = qNW_D = \sqrt{2q\varepsilon N V_{bi}} \tag{4.15}$$

and

$$V_{bi} = \phi_{B0} - \Delta\phi_{B,dipole} - \Delta\phi_{B,BGN} - \frac{E_g}{2} + \frac{kT}{q} * \ln\left(\frac{N}{n_i}\right) \tag{4.16}$$

Solving Equations (4.13) - (4.16) requires an iterative approach, whereby a user-defined value for $\Delta\phi_{B,dipole}$ is used to compute E, which is then used to compute $\Delta\phi_{B,image}$ and $\Delta\phi_{B,dipole}$. The self-consistent solution is reached when the user-defined and computed values for $\Delta\phi_{B,dipole}$ match. $\Delta\phi_{B,BGN}$ (SBL due to BGN) is included in Equation (4.16) and is, for simplicity, taken as a best-fit curve to TCAD BGN modeling results:

$$\Delta\phi_{B,BGN} = \exp\left[-17.371 + 0.323 * \ln(N)\right] \tag{4.17}$$

All that is left now is to implement Equations (4.13) – (4.17) and to generate a best-fit curve

for $\Delta\phi_B$ *vs*. *E*. It is noteworthy that *E* in Equation (4.5) does not account for SBL and so is not the actual *E*, which would be lower as ϕ_B is reduced when *N* increases. Thus, the "trick" here is to develop a simple relationship between the self-consistent solution for SBL and the first-order solution for *E*. In implementing the SBL model, however, realistic values for β and λ must be used in order to give SBL results that are representative of actual contacting materials (*e.g.* NiSi, PtSi). This is where the data presented in [1] becomes particularly useful, as it provides a very good starting point for extracting realistic values for β and λ. Fig. 4.1 shows the overlay of the experimental data in [1] to TCAD modeling curves for ρ_c *vs*. *N* for contacts to n-type and p-type Si, as well as ITRS 2007 specifications for ρ_C [3].

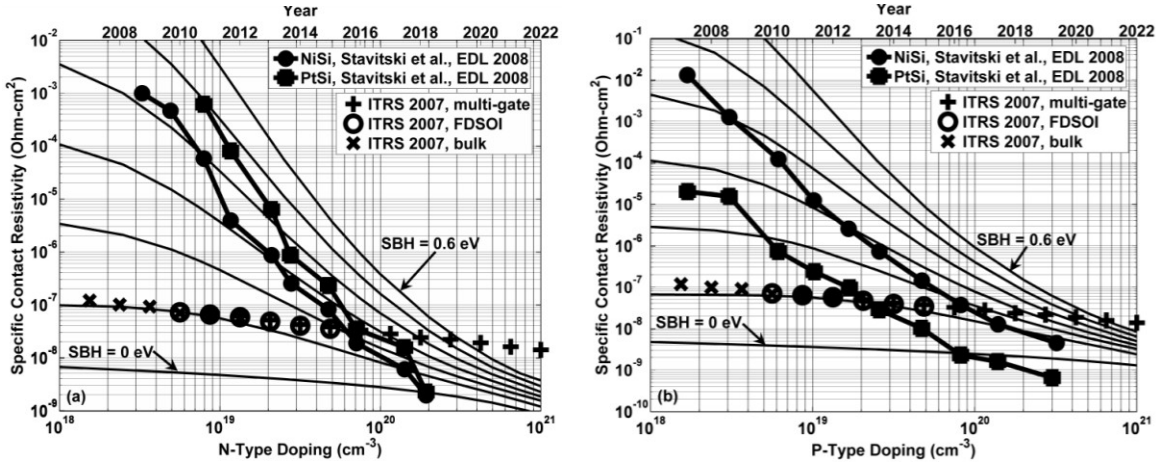

Fig. 4.1. ρ_c vs. doping concentration for NiSi and PtSi contacts to (a) n-type Si and (b) p-type Si (experimental data extracted from [1]). Also shown are modeling curves (solid lines, ϕ_B = 0-0.6 eV, $\Delta\phi_B$ = 0.1 eV) and ITRS 2007 specifications, with the ITRS timeline shown as the top x-axis.

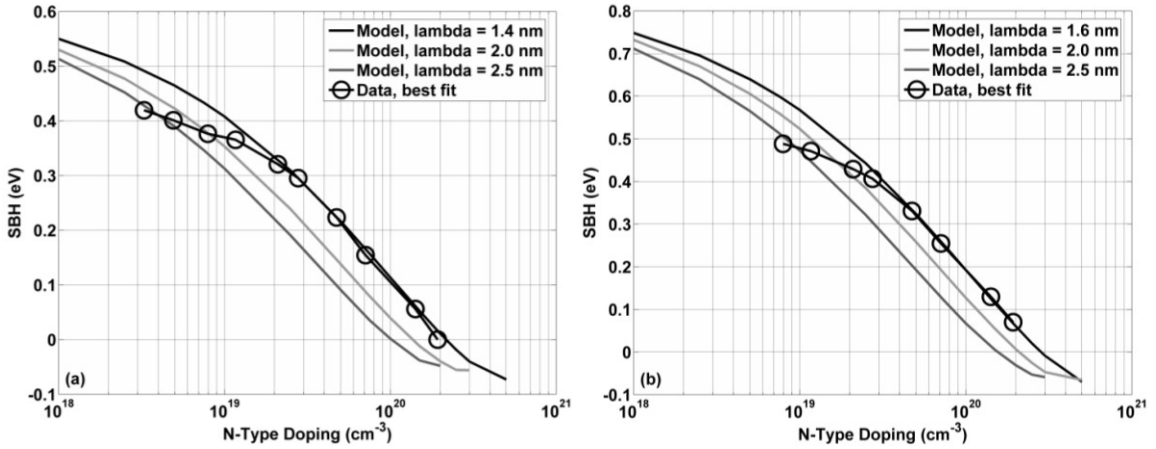

Fig. 4.2. ϕ_B vs. *N* for (a) NiSi and (b) PtSi contact to n-type Si, with the best-fit data curve compared to the modeled curves for different values of λ. β = 1 in all cases.

From Fig. 4.1, ϕ_B *vs*. *N* is extracted by interpolating between the modeled curves over the range of values for *N*. Figs. 4.2 and 4.3 compare the derived ϕ_B *vs*. *N* data for contacts to n-type and p-type Si, respectively, against the self-consistent SBL model for varying values of λ and β. It is important to point out that the slope of the ϕ_B *vs*. *N* curve is affected only by β. A change in

λ merely shifts the curve to the left or right. This makes it relatively straightforward to extract values for λ and β. For contact to n-type Si, $\beta = 1$ and $\lambda = 1.4$ nm and 1.6 nm for NiSi and PtSi ($\phi_{B0} = 0.65$ eV and 0.87 eV), respectively. For contact to p-type Si, $\beta = 0.75$ and $\lambda = 0.75$ nm and 1 nm for NiSi and PtSi ($\phi_{B0} = 0.47$ eV and 0.25 eV), respectively.

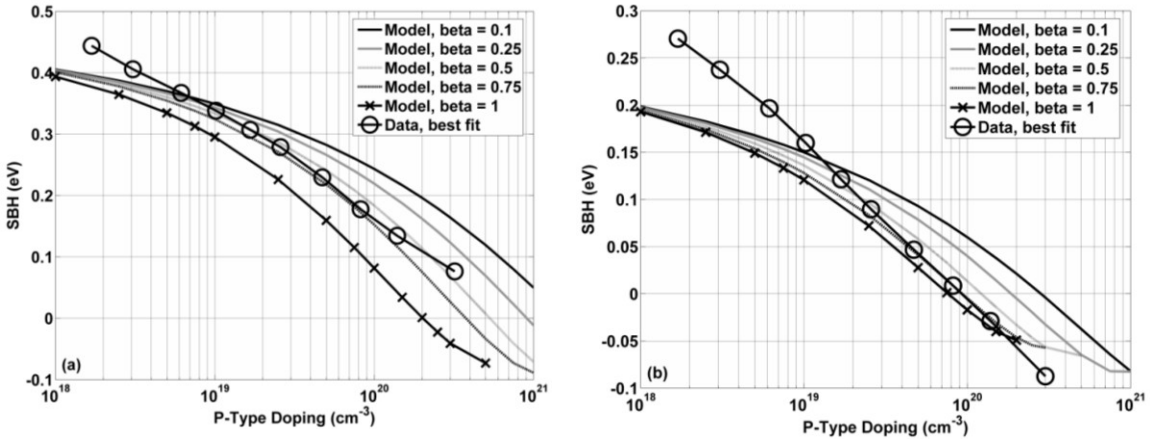

Fig. 4.3. ϕ_B vs. N for (a) NiSi and (b) PtSi contact to p-type Si, with the best-fit data curve compared to the modeling curves for different values of β. For NiSi, $\lambda = 0.75$ nm, while for PtSi, $\lambda = 1$ nm.

Close examination of Figs. 4.2 and 4.3 reveals the difficulty of developing a fully predictive SBL model. Particularly interesting in Figs. 4.2(a) and (b) is the fact that λ starts off at a large value (2.5 nm or higher) for $N < 1\text{x}10^{19}$ cm^{-3}. As N increases, λ drops and eventually reaches a lower limit, as Fig. 4.4 shows. In other words, for low N, the self-consistent SBL model underestimates SBL. There are two possible explanations for this. First is the possibility that ϕ_{B0} used in the model is too high. This was investigated, and lowering ϕ_{B0} still resulted in the best-fit data curve diverging from the modeled curve at low N and constant λ. This points to the second and more likely possibility that λ is a function of N. In [12], λ was noted to be a function of E (and, by extension, N); however the relationship given in [12] was such that λ increased with E, which is opposite to what is observed here. As mentioned previously, λ is the average Heine tail or electron wavefunction decay length into the semiconductor, which is the distance from the Schottky contact at which the amplitude of the electron wavefunction penetrating into the semiconductor drops by the factor $1/e$. This decay length is a function of energy, since the metal-induced gap states (MIGS) induced by the penetrating wavefunctions results in an energetic distribution of donor- and acceptor-like states within the bandgap. The energy level at which the concentration of these states is equal (and to which the Fermi level E_F is pinned) is called the branch point (E_B) [13], and is also referred to as the charge neutrality level (E_{CNL}) [14]. For NiSi and PtSi, the hole SBH is smaller than the electron SBH, because there are more acceptor-like MIGS near to the conduction band than there are donor-like MIGS near to the valence band so that E_B is located closer to the valence band.

For a contact to n-type Si, as N increases, W_D drops and approaches λ, so that the energy-integrated charge density due to acceptor-like MIGS drops. (MIGS arises from electrons that try to tunnel through the semiconductor but cannot, due to a barrier width that is too large, so that the electron wavefunction simply decays into the semiconductor; as W_D drops and more electrons can tunnel through, the "charge" within the bandgap generated by the energetic and spatial distributions of these wavefunctions drops due to the finite tunnel barrier width, limiting

wavefunction penetration into the semiconductor over the energy range of qV_{bi}.) At the same time, ϕ_B drops due to increasing E, which not only reduces W_D even further but also reduces the energy range over which W_D can "contain" MIGS (*i.e.*, $qV_{bi} << E_G$). (These two effects compete against each other, the first to reduce the energy-averaged λ and the second to restore λ to some value, as N increases. The reason why λ does not increase again at very high values of N could be because the ionized donor atoms within the depletion region screen the acceptor-like MIGS, so that λ levels off at high N.)

It is very possible that the extracted λ values shown in Fig. 4.4 are dependent on fabrication process conditions, such as silicidation temperature and time. In [15], for example, it was shown that a Ni diffusion tail extends from the NiSi layer, and that this tail increases with the thermal budget, unless passivating species such as F or N are used [16]. Such a metal diffusion tail would effectively increase λ due to trap-assisted tunneling, so that λ is a function of not only N but also the silicidation conditions (including pre-silicidation implantation of F or N, which would be expected to reduce λ).

The ϕ_B *vs*. N curves for contacts to p-type Si in Fig. 4.3 show that λ is smaller, and that it does not approach a lower limit at high N, in contrast to the curves for contacts to n-type Si. This could be due to differences in the energetic distributions of donor- *vs*. acceptor-like MIGS. It is also possible that the aforementioned metal diffusion tails extending from the silicide may be a function of N. In other words, it is possible that dopants also have a passivating effect (although less effective than F or N) on the silicide, either at the interface or in the silicide grain boundaries, and that B may be a more effective passivation species than As.

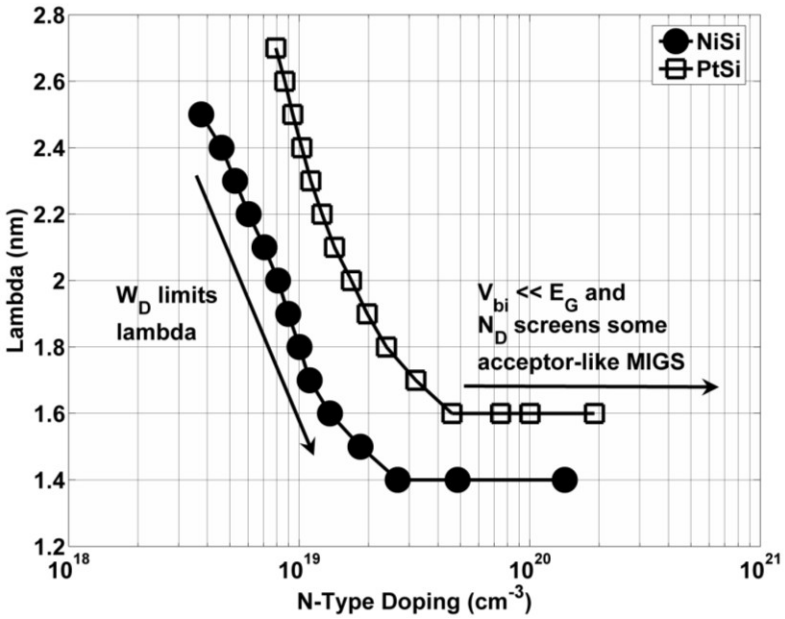

Fig. 4.4. λ vs. N for NiSi and PtSi contact to n-type Si. λ values were extracted from the intersections of the modeled curves and experimental data curves.

Also, for contacts to p-type Si, $\beta < 1$ so that SBL is overestimated at low N. This is consistent with the fact that, for the fabricated structures in [1], the heavily-doped regions were annealed prior to forming NiSi or PtSi. This means that the B atoms were mostly situated in substitutional sites before the contacts were formed. It was shown in [17] that, for p-type contacts in particular,

this reduces the efficiency of B segregation at the silicidation front, due to the extra energy needed to break the B-Si bond. As a result, many B dopants end up as substitutional dopants within the NiSi or PtSi layer, counteracting the substitutional B atoms on the Si side of the Schottky junction for SBL [17], [18]. Thus, it is entirely possible that p-type contacts formed by an alternative method, such as an implant-to-silicide (ITS) process [17], could achieve $\beta \sim 1$ due to the increased ratio of interstitial-to-substitutional B within the silicide.

What this all means is that predictive modeling of SBL is extremely difficult, since both β and λ seem to rely heavily upon the fabrication process conditions. As a result, the closed-form SBL model developed here is only claimed to be representative of the process conditions in [1] for N > mid-10^{19} cm^{-3}. Using the extracted values for β and λ, ϕ_B *vs*. E (units: V/cm) can now be expressed as a 3rd-order polynomial (with a coefficient of determination, or R^2, fit greater than 0.999 for all cases here). Equations (4.18) – (4.21) respectively show this relationship for NiSi to n-type Si, PtSi to n-type Si, NiSi to p-type Si, and PtSi to p-type Si. These equations are then used to obtain $\phi_{B,avg}$ by replacing E with ($E \pm \sigma_E$) and taking the average over $\pm 1\sigma$, as Equation (4.22) shows.

$$\phi_{Bn,NiSi} = \left(-3.6609x10^{-22}\right)E^3 + \left(1.2908x10^{-14}\right)E^2 - \left(1.602x10^{-7}\right)E + 0.6092 \tag{4.18}$$

$$\phi_{Bn,PtSi} = \left(-3.4406x10^{-22}\right)E^3 + \left(1.2815x10^{-14}\right)E^2 - \left(1.7903x10^{-7}\right)E + 0.8278 \tag{4.19}$$

$$\phi_{Bp,NiSi} = \left(-1.1742x10^{-22}\right)E^3 + \left(5.2518x10^{-15}\right)E^2 - \left(8.7919x10^{-8}\right)E + 0.4237 \tag{4.20}$$

$$\phi_{Bp,PtSi} = \left(-8.5407x10^{-22}\right)E^3 + \left(1.7094x10^{-14}\right)E^2 - \left(1.1853x10^{-7}\right)E + 0.2172 \tag{4.21}$$

$$\phi_{B,avg} = \frac{\left[\phi_B\left(E + \sigma_E\right) + \phi_B\left(E - \sigma_E\right)\right]}{2} \tag{4.22}$$

$\phi_{B,avg}$ is used to find $\rho_c \pm \sigma_{\rho c}$, as well as σ_{E00}; however, ϕ_{B0} is used to find W_D and therefore E (which is also used to find σ_{E00}) and σ_E. This creates a problem, in that no closed-form solution exists, since σ_E affects $\sigma_{\phi B}$, which affects σ_E and σ_{E00}. Thus the average ρ_c is overestimated for small contact areas, since ϕ_{B0} is used to find σ_E. To largely circumvent this problem, ϕ_{B0} is only used to find W_D and therefore ($E \pm \sigma_E$) for Equations (4.18)-(4.22). For everything else, separate values of W_D and ($E \pm \sigma_E$) are calculated using the results of Equation (4.22), which is used to find ($E_{00} \pm \sigma_{E00}$) and therefore ($\rho_c \pm \sigma_{\rho c}$). This results in a semi-closed-form solution, where the only discrepancy remaining is that $\sigma_{\phi B}$ depends on ϕ_{B0}. This results in some overestimation of $\sigma_{\phi B}$ and therefore $\phi_{B,avg}$ for small contact areas.

One final point to emphasize with regard to the analytical model presented here is that it has its roots in the WKB approximation. Thus, SB tunneling will be overestimated, meaning that ρ_c is underestimated. This can be corrected somewhat by using larger values for m^* than in the TCAD modeling approach in Section II. Such an approach was shown to be effective in the low-E regime [8]; however, for high N, E is very large and so the m^* required to "tune" the analytical model in order to reasonably fit the TCAD results in Fig. 4.1 may vary with ϕ_B. Thus, a pair of best-fit expressions for the "tuned" m^* for electrons and holes are shown, respectively, in

Equations (4.23) and (4.24), as a function of ϕ_B. Using these values for m^*, a very close fit (Fig. 4.5) between the analytical model and the TCAD results in Fig. 4.1 is achieved. Now the effect of RDF on ρ_c can be analyzed with confidence in the accuracy of the analytical model.

$$m^*_{n,tuned} = 0.3872 * \phi_{Bn}^{-0.7719} \tag{4.23}$$

$$m^*_{p,tuned} = 0.3662 * \phi_{Bp}^{-1.1912} \tag{4.24}$$

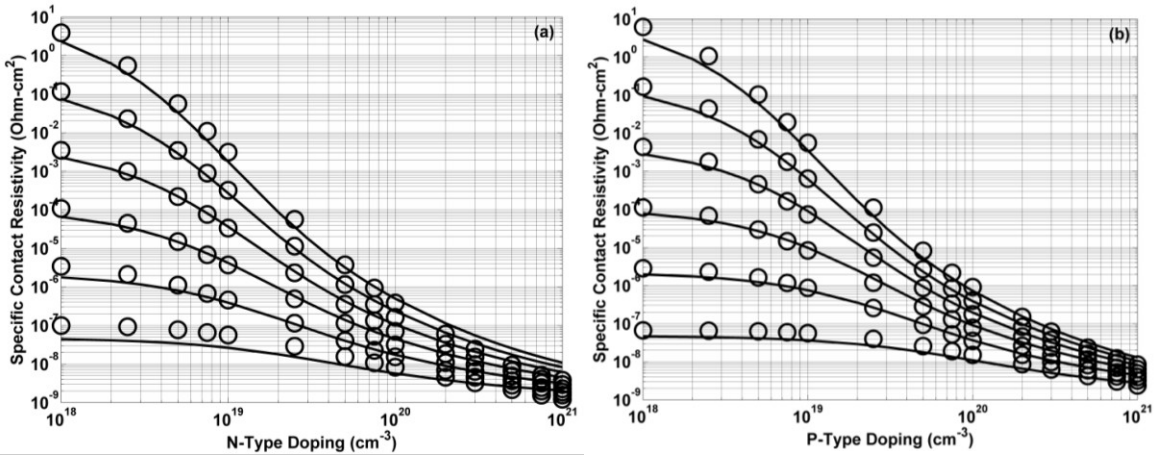

Fig. 4.5. Analytical (solid lines) *vs.* TCAD (circles) comparison of ρ_c *vs.* N for (a) n-type contacts and (b) p-type contacts. ϕ_B is varied from 0.1 eV to 0.6 eV, in 0.1 eV increments, and SBL is excluded.

4.4 Modeling Results

Fig. 4.6 shows the analytical modeling results for $\rho_{c,avg}$ and $\phi_{B,avg}$ *vs.* contact area, for NiSi and PtSi contacts to n-type Si. Similar plots for contacts to p-type Si are shown in Fig. 4.7. It is noted that $\rho_{c,avg}$ in Fig. 4.7(b) for PtSi at $N = 2\text{x}10^{20}$ cm^{-3} is solvable only for contact areas ≤ 3 nm^2. This is because, for larger contact areas, $\phi_{B,avg}$ is negative and so E is imaginary. This is a limitation of the analytical model, in that if N is too large, a zero or slightly negative ϕ_B results and does not permit a solution.

As expected, an increase in N reduces $\rho_{c,avg}$, but more interestingly the contact area below which $\rho_{c,avg}$ increases significantly (~ 10-20 nm^2) is largely independent of N and the contacting material (*i.e.*, NiSi *vs.* PtSi). This has implications in particular for nanowire MOSFETs with dopant-segregated Schottky (DSS) source/drain structure to minimize source/drain series resistance, as in [19]: N must increase (or ϕ_{B0} must drop) in order to achieve the same $\rho_{c,avg}$ as the nanowire diameter is reduced. The situation is actually worse than this, not only for nanowire MOSFETs but also for other ultra-thin body MOSFETs, since this model excludes quantum confinement (QC) effects which would be very significant for nanowire diameters below 6 nm (contact area < 28.3 nm^2) [20]-[22].

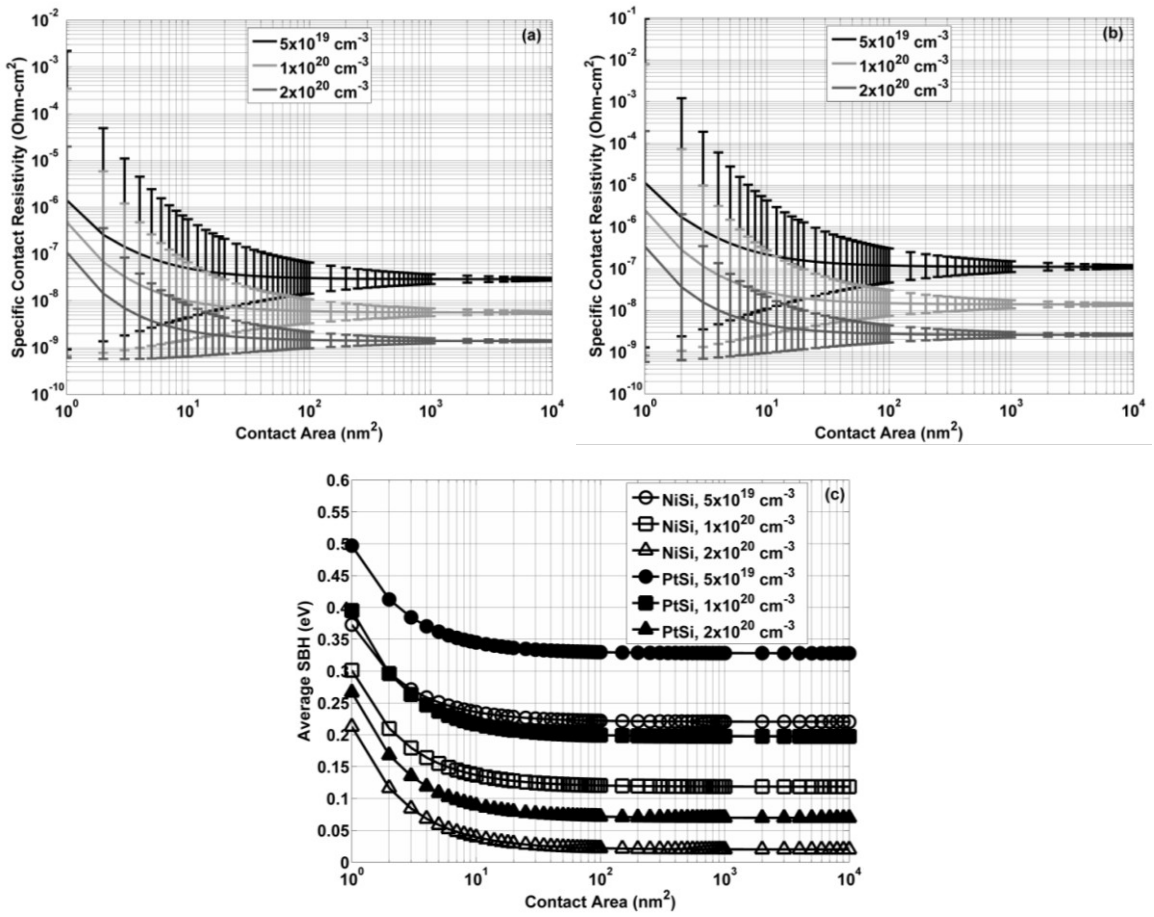

Fig. 4.6. Analytical model results for n-type $\rho_{c,avg}$ vs. contact area for (a) NiSi and (b) PtSi, as well as (c) average ϕ_{Bn} vs. contact area for NiSi and PtSi contacts.

Considering that the increase in $\rho_{c,avg}$ due to RDF is only significant in the contact-area regime where QC is also significant, it would be fair at this point to consider QC, at least to first order. Since QC effects are more significant in nanowire FETs than planar ultra-thin body MOSFETs [21], the nanowire MOSFET is considered here as the worst case. For an infinite quantum well, the change in ϕ_{B0} due to QC is (in electron-volts) $0.376/m^*d^2$, where d is the size of the quantum well. ($m^* = 0.4^*m_0$ gives a reasonable fit to the results in [20]-[22] for Si nanowires.) In terms of DSS nanowire MOSFET contact area, this can be re-expressed as $(0.376\pi)/(4m^*A)$. This shift in ϕ_B must be included in the self-consistent SBL model, meaning that a different self-consistent SBL solution exists for each value of contact area in the QC regime. In other words, closed-form best-fit SBL equations, such as Equations (4.18)-(4.21), would change with contact area, leading to a matrix of equations for each combination of values for dopant concentration and contact area. For simplicity, here the effect of QC is considered independently of RDF, so that QC and RDF can be compared in terms of their effect on ρ_c for small contact areas. The results are shown in Fig. 4.8 for NiSi contacts to n-type Si, which show that, even for the worst-case QC, the effect of RDF on ρ_c is generally stronger, and more so as N increases.

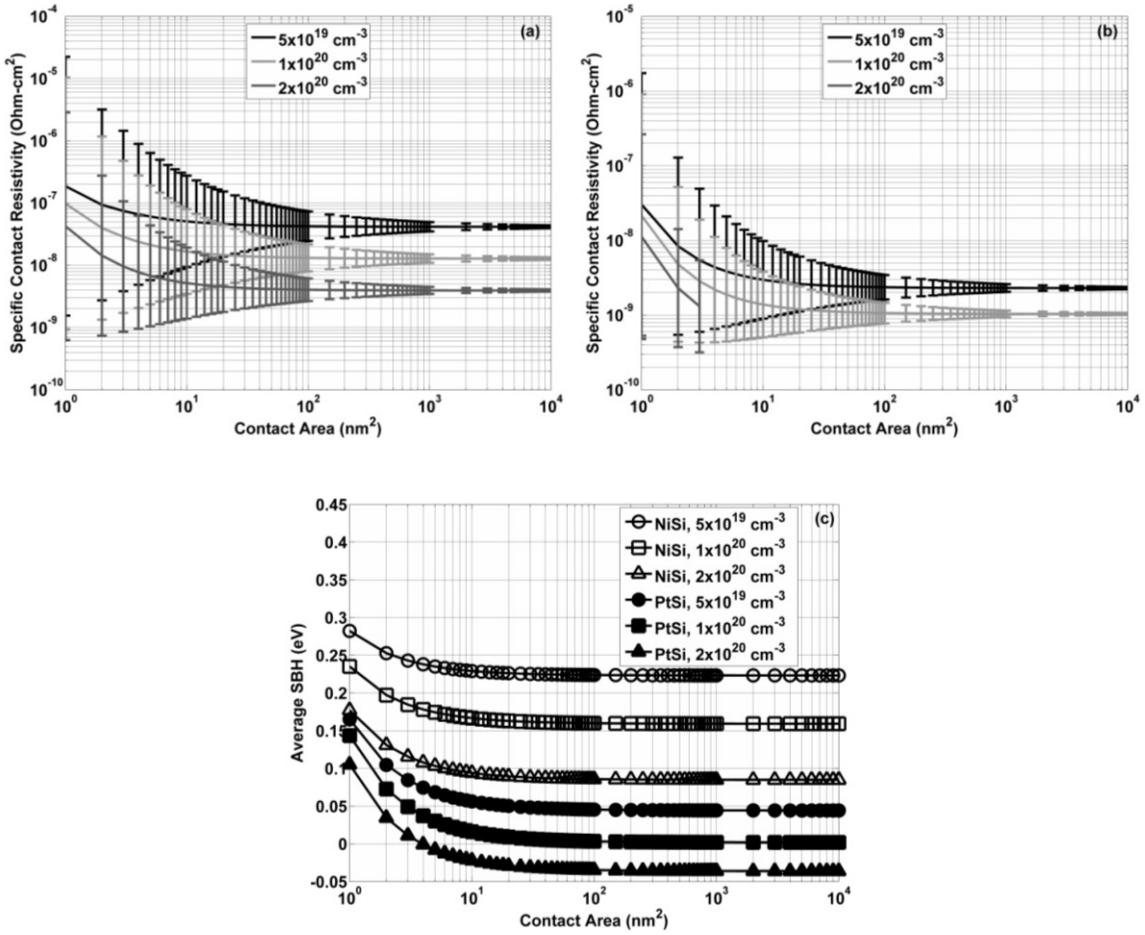

Fig. 4.7. Analytical model results for p-type $\rho_{c,avg}$ vs. contact area for (a) NiSi and (b) PtSi, as well as (c) average ϕ_{Bp} vs. contact area for NiSi and PtSi contacts.

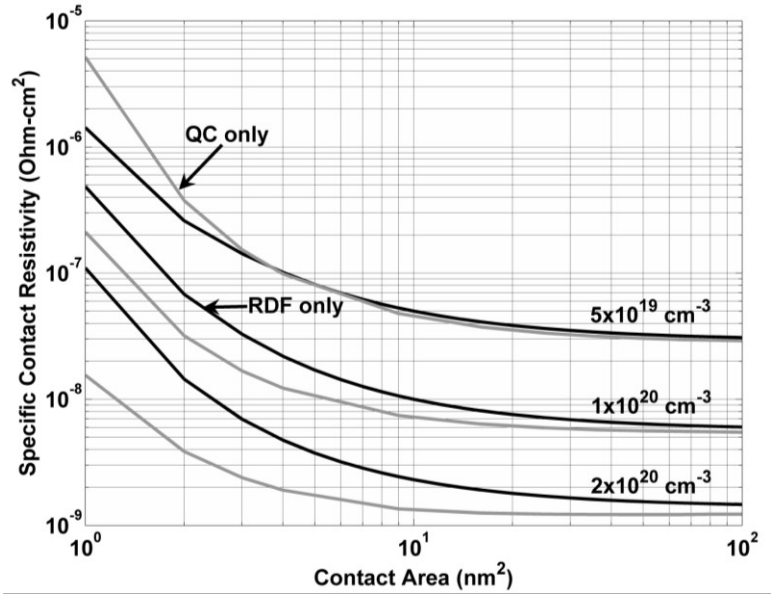

Fig. 4.8. Analytical model results for ρ_c vs. nanowire contact area for NiSi contacts to n-type Si, comparing the individual effects of RDF and QC.

Thus, RDF will dominate ρ_c variation (and the increase in $\rho_{c,avg}$) for contact areas at and near the end of the CMOS technology roadmap (20-30 nm^2 for FinFETs, considering the modeling results in [9] and assuming a 3:1 fin aspect ratio, and possibly smaller for nanowires, depending on nanowire diameter scalability). The spread in ρ_c due to RDF for such small contact areas is significant and grows to well above 1 decade for contact areas less than 10 nm^2, as the error bars in Figs. 4.6 and 4.7 show, but is reduced as N increases. For $N > 1x10^{20}$ cm^{-3} and for contact areas ~ 20 nm^2, $\rho_c + \sigma_{\rho c} \sim 1x10^{-8}$ Ohm-cm^2 and $\rho_{c,avg} < 1x10^{-8}$ Ohm-cm^2, which exceeds end-of-roadmap ITRS requirements, regardless of whether NiSi or PtSi is used as the contacting material.

From this analysis, it would seem that the effect of RDF on ρ_c will not be a limiting factor in transistor scaling, since the spread in ρ_c can be reduced by increasing N, which also reduces $\rho_{c,avg}$.

4.5 Summary

An analytical model has been developed to describe the effect of RDF on ρ_c for very small contact area. This was achieved by extracting the dependence of ρ_c on electric field at the contact interface through a characteristic energy term in the ρ_c model. The variation in this electric field with doping is easily derived to a first order; however, variations in contact SBH and therefore ρ_c due to the effect of RDF on SBL require a self-consistent solution that is non-trivial and likely dependent on process conditions. To circumvent this problem, a semi-closed-form solution is achieved by developing a set of best-fit polynomial equations to the self-consistent SBL solutions, which are calibrated to empirical data. The analytical ρ_c model is then calibrated to TCAD ρ_c simulations accounting for thermal and tunneling current at the contact, to result in accurate ρ_c prediction over a large range of doping levels. The resulting model predicts that ρ_c variation due to RDF drops as the doping concentration increases. Only for contact areas less than 10 nm^2 does the average ρ_c increase significantly due to RDF; however, this effect can be offset with a modest increase in doping. Quantum confinement will also increase ρ_c for sub-10 nm^2 contact areas, but this effect is not as significant as RDF. According to the model, an active dopant concentration of $2x10^{20}$ cm^{-3}, which is readily achievable with modern doping processes, is more than sufficient to achieve low ρ_c and ρ_c variation at ultimately small scales, even for large zero-field barrier contacts, *e.g.* for PtSi contact to n-type Si.

4.6 References

[1] N. Stavitski, M. J. H. van Dal, A. Lauwers, C. Vrancken, A. Y. Kovalign, R. A. M. Wolters, "Systematic TLM measurements of NiSi and PtSi specific contact resistance to n- and p-type Si in a broad doping range," *IEEE Elec. Dev. Lett.*, vol. 29, no. 4, pp. 378-381, Apr. 2008.

[2] N. Stavitski, M. J. H. van Dal, A. Lauwers, C. Vrancken, A. Y. Kovalign, R. A. M. Wolters, "Evaluation of transmission line model structures for silicide-to-silicon specific contact resistance extraction," *IEEE Trans. Elec. Dev.*, vol. 55, no. 5, pp. 1170-1176, May 2008.

[3] *International Technology Roadmap for Semiconductors (ITRS)*. [Online]. Available: http://public.itrs.net

[4] R. A. Vega, T.-J. King Liu, "Three-Dimensional FinFET Source/Drain and Contact Design Optimization Study," *IEEE Trans. Elec. Dev.*, vol. 56. no. 7, pp. 1483-1492, July 2009.

[5] A. Y. C. Yu, "Electron Tunneling and Contact Resistance of Metal-Silicon Contact Barriers," *Solid-State Electronics*, vol. 13, pp. 239-247, 1970.

[6] K. Varahramyan, E. J. Verret, "A Model for Specific Contact Resistance Applicable for Titanium Silicide-Silicon Contacts," *Solid-State Electronics*, vol. 39, no. 11, pp. 1601-1607, 1996.

[7] *User's Manual for Sentaurus Device*, Synopsys Co., Mountainview, CA.

[8] R. A. Vega, T.-J. King Liu, "A Comparative Study of Dopant-Segregated Schottky and Raised Source/Drain Double-Gate MOSFETs," *IEEE Trans. Elec. Dev.*, vol. 55, no. 10, pp. 2665-2677, Oct. 2008.

[9] R. A. Vega, K. Liu, T.-J. King Liu, "Dopant-Segregated Schottky Source/Drain Double-Gate MOSFET Design in the Direct Source-to-Drain Tunneling Regime," *IEEE Trans. Elec. Dev.*, (to be published, Sept. 2009).

[10] F. A. Padovani, R. Stratton, "Field and Thermionic-Field Emission in Schottky Barriers," *Solid State Electronics*, vol. 9, pp. 695-707, 1966.

[11] Y. Taur, T. H. Ning, "Fundamentals of Modern VLSI Devices," *Cambridge University Press*, pp. 200-202.

[12] K. Shenai, E. Sangiorgi, K. C. Saraswat, R. M. Swanson, R. W. Dutton, "Accurate Barrier Modeling of Metal and Silicide Contacts," *IEEE Elec. Dev. Lett.*, vol. 5, no. 5, pp. 145-147, May 1984.

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

[14] V. N. Brudnyi, S. N. Grinyaev, V. E. Stepanov, "Local neutrality conception: Fermi level pinning in defective semiconductors," *Physica B*, vol. 212, pp. 429-435, Dec. 1994.

[15] M. Tsuchiaki, K. Ohuchi, C. Hongo, "Junction Leakage Generation by NiSi Thermal Instability Characterized Using Damage-Free n+/p Silicon Diodes," *Jpn. J. Appl. Phys.*, vol. 43, no. 8A, pp. 5166-5173, 2004.

[16] M. Tsuchiaki, A. Nishiyama, "Substrate Orientation Dependent Suppression of NiSi Induced Junction Leakage by Fluorine and Nitrogen Incorporation," *Jpn. J. Appl. Phys.*, vol. 47, no. 4, pp. 2388-2397, 2008.

[17] T. Yamauchi, Y. Nishi, Y. Tsuchiya, A. Kinoshita, J. Koga, K. Kato, "Novel doping technology for a 1 nm NiSi/Si junction with dipoles comforting Schottky (DCS) barrier," *IEDM Tech. Dig.*, pp. 963-966, 2007.

[18] T. Yamauchi, A. Kinoshita, Y. Tsuchiya, J. Koga, K. Kato, "1 nm NiSi/Si junction design based on first-principles calculation for ultimately low contact resistance," *IEDM Tech. Dig.*, pp. 385-388, 2006.

[19] E. J. Tan, K. L. Pey, N. Singh, G. Q. Lo, D. Z. Chi, Y. K. Chin, L. J. Tang, P. S. Lee, C. K. F. Ho, "Nickel-Silicided Schottky Junction CMOS Transistors with Gate-All-Around Nanowire Channels," *IEEE Elec. Dev. Lett.*, vol. 29, no. 8, pp. 902-905, Aug. 2008.

[20] C. Harris, E. P. O'Reilly, "Nature of the bandgap of silicon and germanium nanowires," *Physica E*, vol. 32, pp. 341-345, 2006.

[21] B. Delley, E. F. Steigmeier, "Size dependence of band gaps in silicon nanostructures," *Appl. Phys. Lett.*, vol. 67, no. 16, pp. 2370-2372, Oct. 1995.

[22] D. D. D. Ma, C. S. Lee, F. C. K. Au, S. Y. Tong, S. T. Lee, "Small-Diameter Silicon Nanowire Surfaces," *Science*, vol. 299, pp. 1874-1877, Mar. 2003.

Chapter 5

High-k Trench Isolation as an Alternative to FinFETs for Ultimate Scalability

5.1 Introduction

Multi-gate MOSFET (MuGFET) structures are more scalable than the conventional planar bulk MOSFET. Examples of MuGFET structures include the double-gate FinFET [1], silicon-on-insulator (SOI) Tri-Gate (TG) MOSFET [2], Omega FET [3], and gate all around (GAA) MOSFET [4]. The disadvantage with MuGFETs is that they are generally more difficult to fabricate due to their three-dimensional (3D) nature. For example, FinFETs require very narrow and tall active regions to suppress short-channel effects (SCE), but the high aspect ratio (AR) makes gate patterning and work function engineering difficult. The bulk TG MOSFET was proposed to circumvent these fabrication challenges [5], since it achieves competitive performance with relaxed active geometry constraints. As with a planar MOSFET, the TG MOSFET fundamentally relies upon steep retrograde (SR) body doping to suppress SCE while also minimizing random dopant fluctuation effects and mobility degradation. This is a problem for low standby power (LSTP) design, because the peak SR doping concentration must increase as the gate length L_G is scaled down, just to suppress thermal leakage. This competes against drain-to-body band-to-band tunneling (BTBT) leakage, which increases with SR doping, to eventually result in zero SR design space for LSTP at small enough scales.

In this study, a modified TG MOSFET design that utilizes a high-permittivity (high-k) dielectric as a channel isolation material for enhanced gate control is proposed. This very shallow high-k trench isolation (HTI) leverages the reverse narrow width effect observed in planar MOSFETs [6], achieving the effect of a gate that physically wraps around more of the active region (as in the FinFET structure). The LSTP performance potential of this new design is analyzed via 3D simulation with Sentaurus Device [7] and is first compared to a conventional bulk TG MOSFET to demonstrate the performance improvement with HTI. It is then directly compared against the FinFET for ultimate LSTP scalability in a realistic gate-pitch and active-pitch-constrained design case.

5.2 Device Structure and Modeling Approach

The single-stripe bulk TG NMOSFET studied here is shown in Fig. 5.1. Note that the active

stripe width $W_{stripe} = L_G$. This is similar to the structure in [5], but with some key differences. First, the SR doping is recessed further into the substrate, so that the peak is below the source/drain extension (SDE) regions; it is referred to herein as the recessed retrograde well (RRW). Since the thickness of the lightly doped body (T_{Si}) is increased, the gate must couple more deeply to this region to suppress thermal leakage. This is the role of the HTI region: it enhances capacitive coupling of the gate to the active sidewalls. Second, deep source/drain (DSD) junctions are eliminated. HTI integration can be similar to that for conventional shallow trench isolation (STI). The HTI surface can be selectively recessed with the aid of ion implantation [8], or the channel can be raised by selective epitaxy [9], so the gate stack wraps around the top portion of the channel.

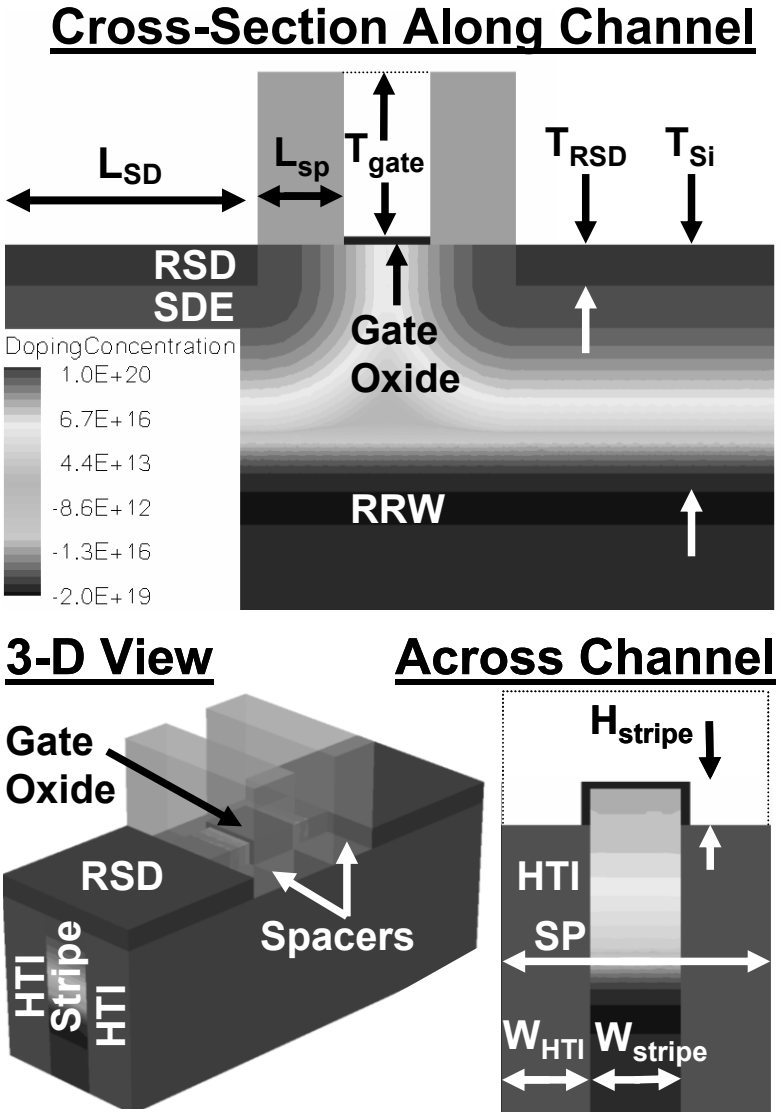

Fig. 5.1. 3D view and 2D cross-sections of the bulk TG MOSFET with HTI modeled in this study. (Sidewall spacers are translucent in the 3D image).

To investigate scaling beyond the perceived limit for the planar bulk MOSFET (L_G = 20 nm) [10], L_G, the nitride gate-sidewall spacer width L_{sp}, and W_{stripe} are each 10 nm, while the gate

oxide thickness t_{ox} = 1 nm. The default values for the widths of the HTI regions (W_{HTI}) are 10 nm (so that the stripe pitch SP = W_{stripe} + 2*W_{HTI} = 30 nm). The gate electrode has a vertical thickness T_{gate} = 20 nm, and it wraps around the top 5 nm of the stripe (*i.e.*, H_{stripe} = 5 nm). A laterally "raised" source/drain (RSD) structure (in which the source/drain contact regions extend over the HTI regions) is used to reduce parasitic resistance and tunneling through the SDE regions, in consideration of the results in [11] and in Chapter 2. The RSD region has a thickness T_{RSD} = 5 nm and is uniformly doped at $1x10^{20}$ cm^{-3}. The length of the source/drain region L_{SD} = 30 nm. The SDE doping extending from the RSD region has a Gaussian profile with a decay length L_{SDE} = 7 nm, corresponding to the distance from the peak where the doping is 1 decade lower ($1x10^{19}$ cm^{-3}). The RRW has a peak concentration of $2x10^{19}$ cm^{-3}, located 30 nm from the top channel surface (*i.e.*, T_{Si} = 30 nm), and a Gaussian profile with decay length L_{RRW} = 3-25 nm. The substrate beneath the RRW is uniformly doped at $1x10^{19}$ cm^{-3}. The HTI is 40 nm tall and reaches a depth of 45 nm from the top channel surface. Note that the HTI is much shallower than the STI that is conventionally used to isolate transistors. The HTI dielectric constant (ε_{HTI}) is varied from 3.9 (SiO_2) to 50 (HfO_2 [12]).

For comparison, a conventional TG (CTG) structure, as in [5], is also modeled. L_G, W_{stripe}, t_{ox}, W_{HTI}, T_{gate}, H_{stripe}, and T_{RSD} are the same as for the HTI TG structure. The body and source/drain doping profiles were optimized through extensive simulation. The SR doping profile has a peak concentration ($2x10^{19}$ cm^{-3}) located 5 nm from the top channel surface, and a Gaussian profile with decay length L_{SR}. The substrate is uniformly doped at $1x10^{18}$ cm^{-3} to minimize BTBT and sub-surface thermal leakage. The gate-sidewall spacers corresponding to the SDE and DSD regions are each 5 nm wide, resulting in a total L_{sp} = 10 nm as for the HTI MOSFET. The SDE and DSD decay lengths are L_{SDE} = 7 nm and L_{DSD} = 18 nm, respectively. The SiO_2 trench isolation is 20 nm tall and reaches a depth of 25 nm from the top channel surface, due to H_{stripe} = 5 nm.

Current flow is modeled using drift-diffusion formulations with doping- and field-dependent mobility models, as in [13], with ohmic contacts. A metallic gate is used (gate leakage is ignored), with ideal threshold voltage tuning through gate work-function engineering. The power supply voltage V_{DD} = 0.75 V and the off-state current I_{OFF} = 24 pA/μm (normalized to W_{stripe} + 2*H_{stripe}), according to ITRS 2007 specifications for L_G = 10 nm [10]. BTBT leakage is modeled using the dynamic non-local BTBT model with default (silicon) parameters [7]. This model creates a non-local mesh only in the regions where BTBT exists, for each bias condition, and can model tunneling in any direction, which is critical for 3D device simulation.

5.3 Conventional Bulk Tri-Gate vs. HTI Tri-Gate MOSFET

Fig. 5.2 shows the leakage floor I_{min} *vs.* L_{SR} (or L_{RRW}) for the CTG (or HTI) structure. For the CTG structure, the high body doping at the drain junction increases BTBT, which prevents I_{min} < I_{OFF}. In other words, there is no design window for L_{SR} to adjust the trade-off between BTBT (which increases with increasing L_{SR}) and SCE (which increases with decreasing L_{SR}). By eliminating the DSD regions and forming a RRW, drain-to-body BTBT is significantly reduced. As ε_{HTI} is increased (*i.e.*, by using HfO_2 rather than SiO_2), the gate fringing field capacitance along the active sidewall increases, reducing sub-surface drain-induced barrier lowering (DIBL), so that I_{min} is further reduced. The minimum I_{min} that can be achieved with HTI is ~ 10x lower than for the CTG structure.

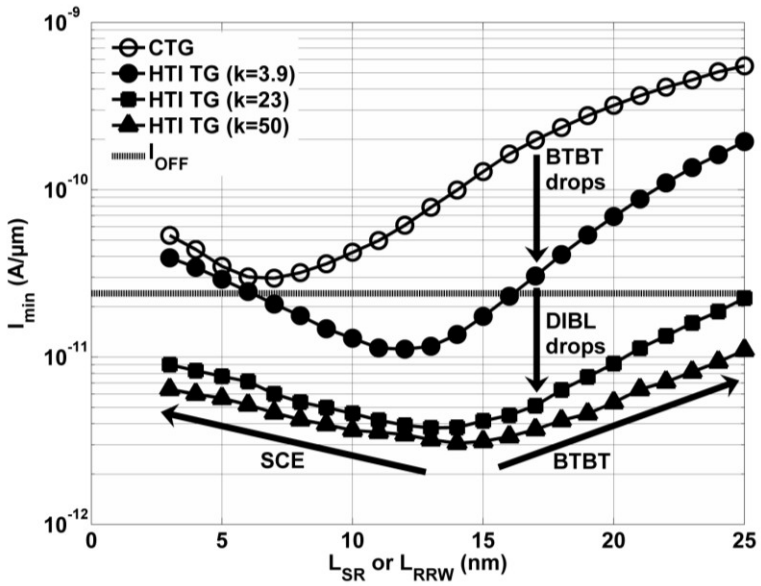

Fig. 5.2. I_{min} *vs.* L_{SR} (CTG structure) and L_{RRW} (HTI MOSFET).

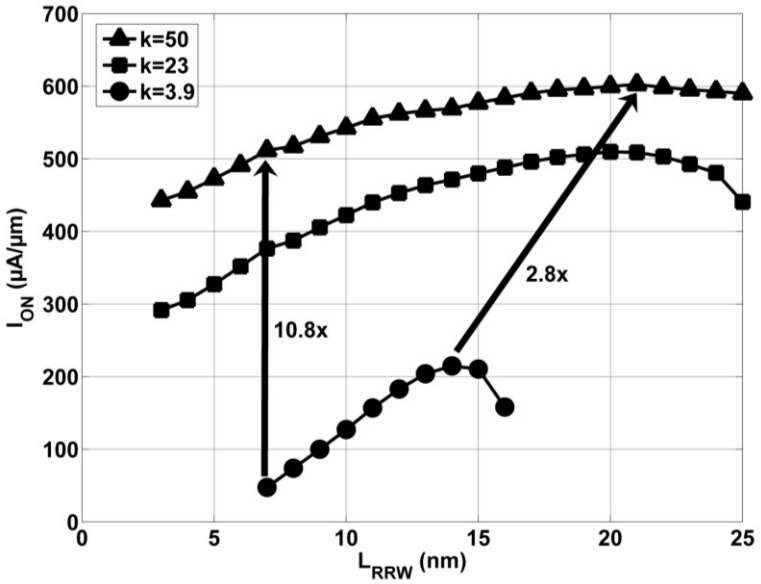

Fig. 5.3. I_{ON} *vs.* L_{RRW} for the HTI MOSFET, with different values of ε_{HTI}.

Fig. 5.3 shows I_{ON} *vs.* L_{RRW} for the HTI MOSFET. While the RRW without DSD reduces I_{min} for SiO_2 isolation (as Fig. 5.2 shows), I_{ON} is low due to sub-surface DIBL. As ε_{HTI} increases, both I_{ON} and the L_{RRW} design space increase. The optimal I_{ON} values are 215, 510, and 603 μA/μm for ε_{HTI} = 3.9, 23, and 50, at L_{RRW} = 14, 20, and 21 nm, respectively. L_{RRW} should be as small as possible, though, to minimize the depletion charge in the channel and therefore performance variation to due random dopant fluctuation [14]. As L_{RRW} is reduced for this purpose, ε_{HTI} must increase to recover I_{ON}.

Simply considering I_{ON} normalized to W_{stripe} + 2*H_{stripe}, as in Fig. 5.3, can be misleading, since it reveals little information about layout area efficiency, especially as W_{HTI} and therefore SP is varied. This is also the case for FinFETs, for which the fin pitch (FP) has a significant effect on

circuit density, but little effect on I_{ON} except at very small scales [15]. Thus, W_{HTI} and therefore SP must be optimized with respect to the trade-off between electrostatic integrity (large W_{HTI}) and layout efficiency (small W_{HTI}). This is shown in Fig. 5.4, where I_{ON} is normalized to SP or $W_{stripe} + 2*H_{stripe}$.

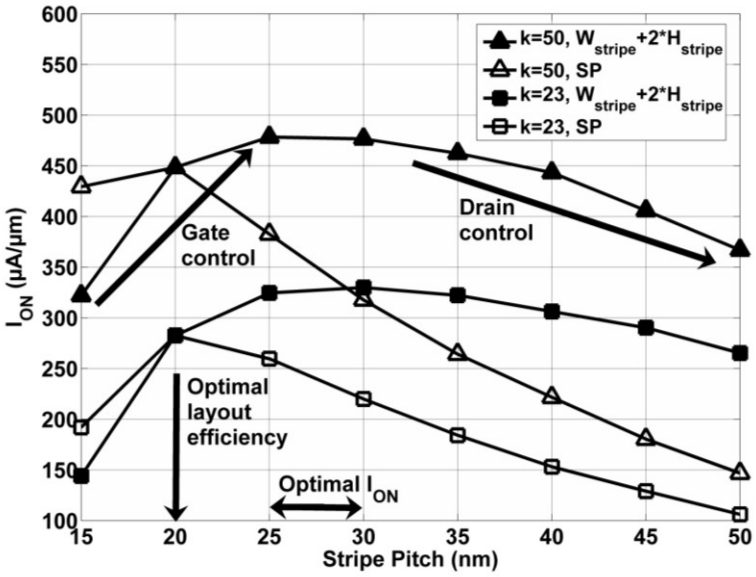

Fig. 5.4. I_{ON} *vs.* SP for the HTI MOSFET with ε_{HTI} = 23 and 50, normalized to SP (open symbols) or W_{stripe} + $2*H_{stripe}$ (closed symbols). L_{RRW} = 5 nm.

In terms of $I_{ON}/(W_{stripe} + 2*H_{stripe})$, the optimal SP ~25-30 nm. As SP increases to this point, the increase in gate fringing fields improves gate control; however, as SP increases further, the drain fringing fields through the HTI regions degrades performance. This degradation is stronger for larger ε_{HTI} and is much like fringe-induced barrier lowering (FIBL) for high-k gate dielectrics [16], except that here it occurs along the active sidewalls. In terms of layout area efficiency (I_{ON}/SP), the optimal SP is smaller, at 20 nm, since increasing SP further gives diminishing returns for I_{ON}. This, in turn, suggests that the optimal W_{HTI} = 5 nm. For large SP and/or low layout density, the HTI region can therefore be applied as a trench liner, with the remaining trench area filled with SiO_2, in order to minimize gate-to-bulk parasitic capacitance.

Admittedly, the modeled I_{ON} falls short of the ITRS specification for LSTP I_{ON} at L_G = 10 nm (935 μA/μm), because only classical drift-diffusion is simulated here and no mobility boost due to strain (which is enhanced at small channel widths [17]) is assumed. It is worthwhile to note that the modeled I_{ON} here is comparable to that of an optimized double-gate MOSFET design (Chapter 2) for the same L_G but at significantly higher V_{DD} (1 V). Thus, it would seem that the bulk HTI MOSFET is competitive with the FinFET for LSTP applications at aggressive scales and so a direct comparison between the two structures is justified.

5.4 FinFET vs. HTI Tri-Gate MOSFET

The HTI and FinFET structures modeled for this comparison are shown in Figs. 5.5(a) and (b), respectively. End-of-roadmap (EOR) ITRS 2007 LSTP specifications are used to define the

boundary conditions for gate pitch (GP = 22 nm), I_{OFF} (29 pA/μm), and V_{DD} (0.7 V). The HTI MOSFET in Fig. 5.5(a) is the same as in Fig. 5.1, except that the total device length is constrained by GP. In other words, $L_G + 2*(L_{SD} + L_{sp})$ = GP. An additional exception is that a dual high-k/low-k spacer is used for both the HTI MOSFET and the FinFET, per the analysis presented in Chapter 3. The length of the outer spacer $L_{sp,lk}$ = 2 nm, while the length of the inner spacer $L_{sp,hk}$ is varied ($L_{sp,lk} + L_{sp,hk} = L_{sp,tot}$). The outer spacer is SiO_2 (ε = 3.9) while the inner spacer is HfO_2 (ε = 23). For the HTI MOSFET, $W_{stripe} = L_G$ and H_{stripe} = 4 nm. The RRW doping, placement, and decay profile, as well as HTI depth, are the same as in Section 5.2. Also, T_{gate} = 15 nm for both the FinFET and HTI MOSFETs.

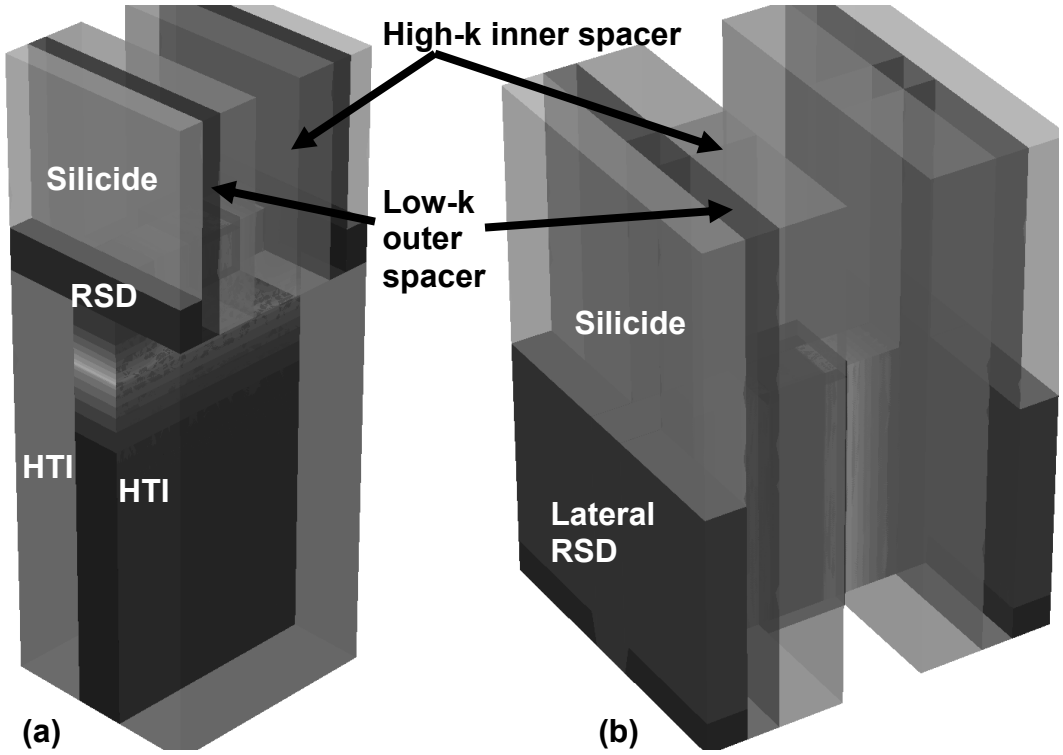

Fig. 5.5. 3D view of the (a) bulk HTI MOSFET and (b) FinFET structures modeled in this study. The HTI regions, sidewall spacers, and silicide contacts are translucent in both images.

For the FinFET in Fig. 5.5(b), the fin width W_{fin} = 3 nm, the fin height H_{fin} = 15 nm, and t_{ox} = 1 nm. The active pitch AP = $W_{fin} + 2*(t_{ox} + T_{gate})$ for the FinFET and $W_{stripe} + 2*W_{HTI}$ for the HTI MOSFET, and is set to 20 nm in both cases. Since LSTP design is considered here, the source/drain and contact structure used for the FinFET is the recessed silicide raised source/drain (RSD) structure discussed in Chapter 2, which minimizes both silicide gating and tunneling through the SDE regions (the source/drain doping is $1x10^{20}$ cm^{-3}, with L_{SDE} varied in the same manner as the HTI MOSFET). Ohmic contacts are assumed between the silicide and silicon, and the silicide has a workfunction of 4.7 eV (assuming NiSi). This silicide region in both the FinFET and HTI MOSFETs represents a combination of both the silicide contact and the contact via which, for such small dimensions, may end up being merged into one structure. The ohmic contact assumption is reasonable, considering the results in Chapter 4 and the fact that the smallest contact area simulated here is 20 nm^2. The density gradient model (DGM) is used to model quantization in the narrow FinFET. The HTI MOSFET is modeled without DGM, in part

due to convergence issues and in part because the I_{DS} vs. V_{GS} curves from the convergent simulations show anomalous behavior. The use of classical carrier computation in the HTI MOSFET is admittedly optimistic, but it is not expected to alter the fundamental conclusion of this study.

5.4.1 Drain Current Normalization in the HTI MOSFET

In Section 5.3, it was mentioned that normalizing I_{DS} with respect to $W + 2*H$ is insufficient, as it reveals little information about layout efficiency. Although normalizing with respect to SP does reveal this information, it is equally insufficient when comparing the HTI MOSFET to a different structure, such as the FinFET. This is because both SP and $W + 2*H$ do not represent the actual physical width of the channel. Therefore, a value for the effective device width W_{eff}, which is representative of the actual channel width, must be derived for the HTI MOSFET. Since it is complex to solve for the additional width of the inversion layer along the active sidewalls controlled by the HTI, a simpler approach is to solve for the gate fringing field capacitance to the active sidewall through the HTI region, or C_{HTI}. The ratio of this capacitance to the gate oxide capacitance (which exists along $W + 2*H$) is then multiplied by W_{HTI} (for W_{HTI} < the HTI depth, such that the gate coupling along the active sidewalls is assumed to stop at a depth W_{HTI} below the HTI surface) to obtain the effective width of the additional channel formed by gate coupling through the HTI region.

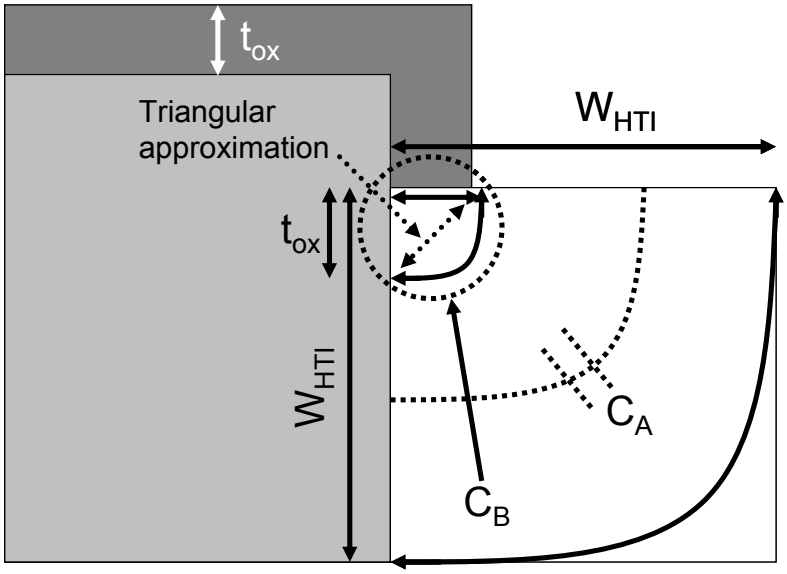

Fig. 5.6. Schematic cross-section of the HTI MOSFET across the channel region, zoomed in to show one side of the device only.

The derivation of C_{HTI} is broken down into two parts: the capacitance along the HTI surface C_A and the capacitance at the gate-oxide-to-HTI junction C_B (Fig. 5.6). To determine C_A, the gate fringing field path through the HTI to the active sidewalls is assumed to follow a quarter-circle. This capacitance is integrated over the "width" of the capacitor, from t_{ox} to W_{HTI}, as Equation (5.1) shows.

$$C_A = \frac{2\varepsilon_{HTI}}{\pi}\int_{t_{ox}}^{W_{HTI}}\frac{dr}{r} = \frac{2\varepsilon_{HTI}}{\pi}\ln\left(\frac{W_{HTI}}{t_{ox}}\right) \tag{5.1}$$

Determining C_B is more difficult and requires some approximation. This is because the width of the capacitive plate along the active sidewall is equal to t_{ox}, but at the gate-oxide-to-HTI junction, the width is effectively zero. Fringing field lines from this corner region are thus linear when running perpendicular to the gate oxide and change to a quarter-circle path running down along the active sidewall. This change in geometry of the fringing field lines suggests that the field strength drops non-linearly along the active sidewall. However, for small t_{ox}, the fringing fields can be assumed as linear over the entire range, thus resulting in a triangular approximation for determining C_B. Assuming the normalized capacitance $C/C_B = 1$ for the shortest path length and $2/\pi$ for the longest path length, C_B is found using Equation (5.2). Now C_{HTI} is found using Equation (5.3), with C_{ox} expressed in Equation (5.4). It is noteworthy that C_{HTI} is expressed over the width of the capacitor, much like how C_{ox} is multiplied by $W + 2*H$, despite the length units cancelling out in the C_{HTI} derivation. Now, the W_{eff} is found using Equation (5.5), which is used herein to normalize I_{ON} in the HTI MOSFET.

$$C_B = \frac{\varepsilon_{HTI}}{t_{ox}} * \frac{t_{ox}}{2} * \left(1 - \frac{2}{\pi}\right) = \frac{\varepsilon_{HTI}}{2}\left(1 - \frac{2}{\pi}\right) \tag{5.2}$$

$$C_{HTI} = C_A + C_B = \frac{\varepsilon_{HTI}}{2}\left(1 - \frac{2}{\pi}\right) + \frac{2\varepsilon_{HTI}}{\pi}\ln\left(\frac{W_{HTI}}{t_{ox}}\right) \tag{5.3}$$

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}\left(W_{stripe} + 2 * H_{stripe}\right) \tag{5.4}$$

$$W_{eff} = W_{stripe} + 2\left(H_{stripe} + W_{HTI}\left[\frac{C_{HTI}}{C_{ox}}\right]\right) \tag{5.5}$$

5.4.2 Pitch-Constrained DC Design Optimization

With GP constrained, co-optimization of L_G, $L_{sp,tot}$, L_{SD}, and L_{SDE} is necessary. An increase in L_G and/or $L_{sp,tot}$ in an effort to increase gate control and therefore I_{ON} will reduce L_{SD} in order to meet the GP boundary condition, which in turn increases source/drain resistance R_{SD} (due to reduced contact area) and may in fact reduce I_{ON}. However, the investigation in Chapter 3 suggests that at extremely small scales, the source/drain contact area is less of a concern than keeping $L_{sp,tot}$ as large as possible to maximize the off-state electrical channel length L_{elec}. In the 2007 ITRS Roadmap, L_G = 8 nm for a half-pitch of 11 nm (GP = 22 nm), meaning the remaining 14 nm must be divided into two sidewall spacers and RSD regions. Assuming a 1 nm resolution limit for L_{SD}, the optimal design for $L_{sp,tot}$ = 6 nm. Fig. 5.7(a) shows the FinFET design surface, while Fig. 5.7(b) shows the HTI design surface. In both cases, L_G and L_{SDE} are co-optimized

with $L_{sp,tot}$ = 6 nm. For the FinFET and HTI MOSFETs, respectively the optimal L_G/L_{SDE} = 5 nm/2 nm and 8 nm/3 nm. Fig. 5.8(a) shows another FinFET design surface, where $L_{sp,tot}$ and L_{SDE} are co-optimized at L_G = 5 nm and demonstrates that, even for L_G < 8 nm, $L_{sp,tot}$ = 6 nm is either optimal or very close to optimal.

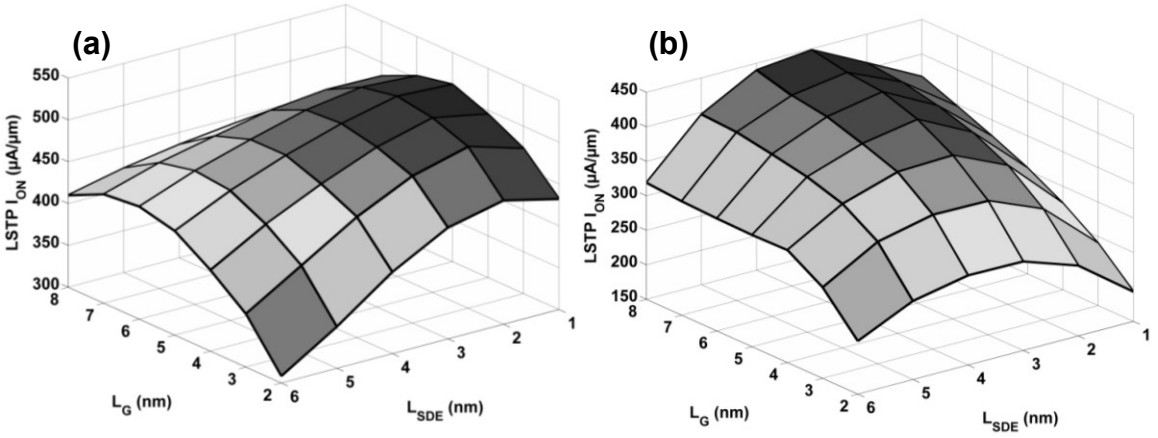

Fig. 5.7. LSTP I_{ON} vs. L_G and L_{SDE} design surfaces for (a) the FinFET and (b) the HTI MOSFET, both with $L_{sp,total}$ = 6 nm.

Making a cut line in Figs. 5.7(a) and (b) at constant L_{SDE}, one finds the optimal I_{ON} vs. L_G for the FinFET and HTI MOSFETs, as Fig. 5.8(b) shows. Although at no design point does I_{ON} in the HTI MOSFET exceed that of the FinFET, the optimal I_{ON} is nonetheless competitive within ~10 % and reveals the utility of HTI in extending bulk LSTP scalability far beyond ITRS projections, which stop at L_G = 20 nm for GP = 64 nm (half-pitch = 32 nm). The HTI MOSFET can be further improved by reducing drain fringing field coupling through the HTI region. This can be achieved by forming a dual trench (DT) structure (henceforth called HTI-DT), whereby the trench dielectric from the high-k/low-k spacer interface outwards, away from the gate electrode, is a low-k material (SiO_2 is used here). The improvement in I_{ON} for small L_G is demonstrated in Fig. 5.8(b).

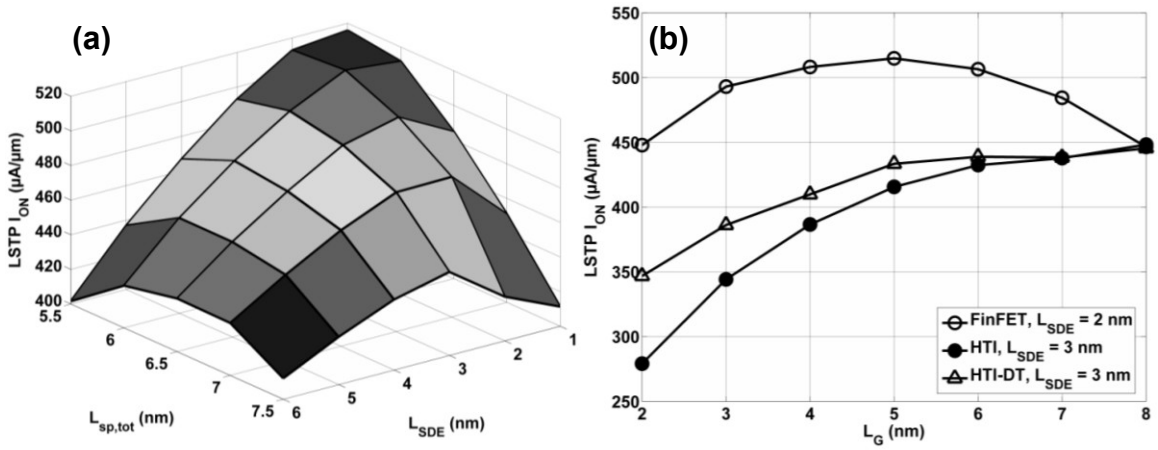

Fig. 5.8. LSTP I_{ON} vs. (a) L_{SDE} and $L_{sp,tot}$ for the FinFET and (b) L_G for the FinFET, HTI, and HTI-DT MOSFETs, with optimal L_{SDE} and $L_{sp,tot}$ = 6 nm. In (a), as L_G is increased, $L_{sp,hk}$ is correspondingly decreased (*i.e.*, the gate electrode displaces the high-k inner spacer).

As Fig. 5.8(b) summarizes, purely from a DC perspective, the FinFET design space for L_G is larger and the optimal L_G is smaller than it is for the HTI MOSFET. From a geometric viewpoint, this means that the optimal L_{SD} is smaller in the HTI MOSFET (1 nm vs. 2.5 nm in the FinFET). This may be an integration challenge concerning the epitaxially grown RSD regions (*i.e.*, loading effects due to carrier gas flow into small regions), the metallic source/drain via regions (due to metal grain size, although metallic carbon nanotube vias are an alternative [18]), and the lithographic alignment tolerance from these vias to the first interconnect level. On the other hand, this conclusion may be premature, because an AC analysis (next sub-section) may result in a smaller optimal L_G, and therefore larger optimal L_{SD}, for either or both the FinFET and HTI MOSFET.

5.4.3 Pitch-Constrained AC Design Optimization

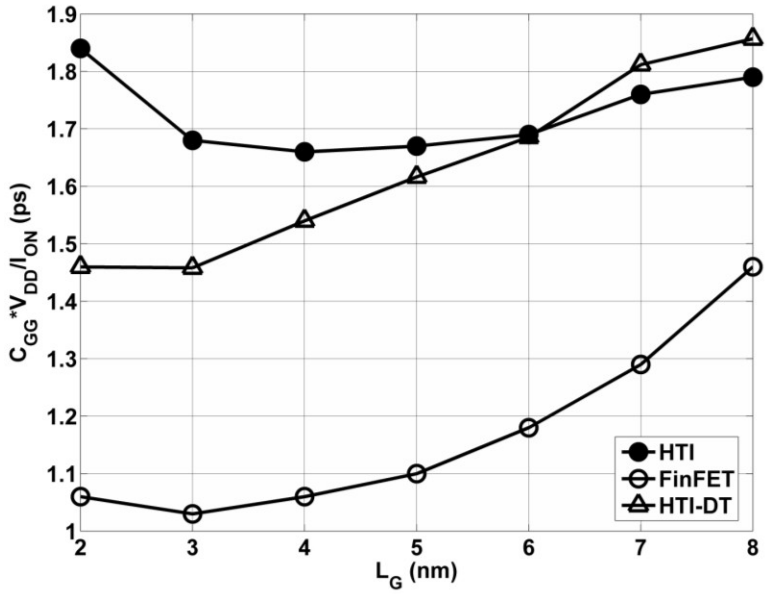

Fig. 5.9. $C_{GG}*V_{DD}/I_{ON}$ for the FinFET, HTI, and HTI-DT MOSFETs, with $L_{sp,tot}$ = 6 nm and L_{SDE} = 2 nm (FinFET) or 3 nm (HTI and HTI-DT).

Using the DC-optimized values for L_{SDE} (2 nm for the FinFET, 3 nm for the HTI and HTI-DT MOSFETs), intrinsic delay $C_{GG}*V_{DD}/I_{ON}$ is simulated from L_G = 2 nm to 8 nm (*i.e.*, C_{GG} is modeled for the design cases in Fig. 5.8(b), to find the intrinsic delay). The results are shown in Fig. 5.9. For the FinFET, the optimal L_G drops from 5 nm (DC optimization) to 3 nm (AC optimization). For the HTI MOSFET, the optimal L_G drops from 8 nm (DC) to 4 nm (AC), while for the HTI-DT MOSFET the AC-optimized L_G = 3 nm. This means the optimal L_{SD} is larger than from the DC optimization, but not because R_{SD} is important. It is instead because C_{GG} scales linearly with L_G. So, even though I_{ON} drops with L_G (in the case of the HTI MOSFET), the reduction in C_{GG} more than offsets this to reduce delay. Eventually, L_G becomes too small and the corresponding drop in I_{ON} more than offsets the reduction in C_{GG}, thus increasing delay. The optimal $C_{GG}*V_{DD}/I_{ON}$ is higher in the HTI (1.66 ps) and HTI-DT (1.46 ps) MOSFETs than in the FinFET (1.03 ps). However, internal loading capacitances (*i.e.*, drain-body capacitance C_{db} and Miller capacitance C_M) will affect the peak overshoot (*i.e.*, Miller

effect) and therefore the inverter switching speed. Thus, a full inverter delay simulation is necessary for a fair comparison between the FinFET and HTI MOSFETs.

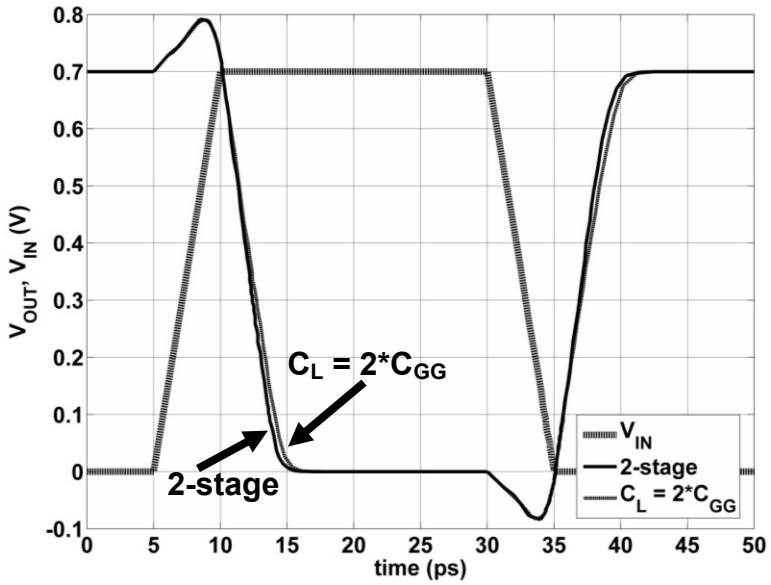

Fig. 5.10. HTI inverter delay simulation results comparing a single stage with $C_L = 2*C_{GG}$ and a 2-stage inverter chain.

Here, an inverter is formed for each structure, with the PMOSFET having the exact same geometry, doping, etc. as the NMOSFET. The PMOSFET width is increased (*i.e.*, the "area" factor in the command file) to achieve a matched inverter. The load capacitance at the output $C_L = 2*C_{GG}$, which approximates the load capacitance of the next inverter stage in a ring oscillator (wiring capacitance is assumed as zero). As Fig. 5.10 shows for the case of the HTI inverter, $C_L = 2*C_{GG}$ is a reasonable approximation, although it results in some overestimation of propagation delay, rise time, and fall time (τ_p, τ_r, and τ_f, respectively). This is because C_{GG} is simulated in the linear regime ($V_{DS} = 0$ V and $V_{GS} = V_{DD}$) where C_{GG} will be highest, but as the inverter switches states, one transistor transitions from cutoff to saturation and then to linear mode, while the other transitions in the opposite direction, resulting in $C_L < 2*C_{GG}$ on average.

As Fig. 5.11 shows, the higher Miller effect in the FinFET offsets the I_{ON} advantage to result in similar τ_p as compared to the HTI-DT inverter (4.54 ps for the FinFET inverter vs. 4.49 ps for the HTI-DT inverter). This points to an interesting property of the HTI and HTI-DT MOSFETs, in that the extra device width that results from gate fringing fields through the HTI regions comes at a smaller cost to C_M than in the FinFET. In the FinFET, adding device width means physically increasing H_{fin} and therefore C_M by the same amount, since the entire flared RSD region overlaps the entire gate sidewall along the sides of the fin. In the HTI and HTI-DT MOSFETs, the RSD region has the same physical height with or without the gate fringing field coupling through the HTI region, but the device benefits from extra gate fringing fields through the HTI regions to effectively increase the device width. There is some gate-to-drain coupling through the HTI regions, though, as evidenced by the reduction in Miller effect with the HTI-DT inverter compared to the HTI inverter, which also has larger $\tau_p = 4.85$ ps, in Fig. 5.11.

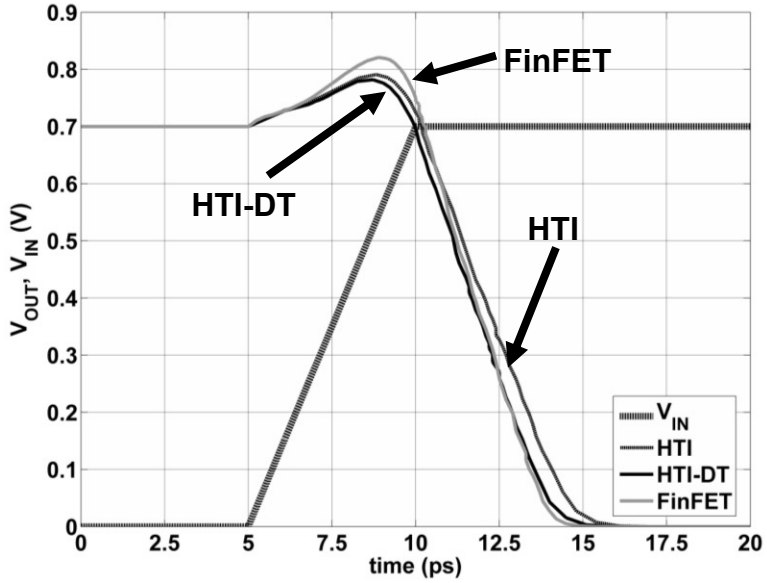

Fig. 5.11. Inverter delay simulation results comparing the HTI, HTI-DT, and FinFET inverters, all with C_L = 2*C_{GG}. Only the pull-down operation is shown here for clarity, since the pull-up operation is symmetric. L_G are the AC-optimized values from Fig. 5.9.

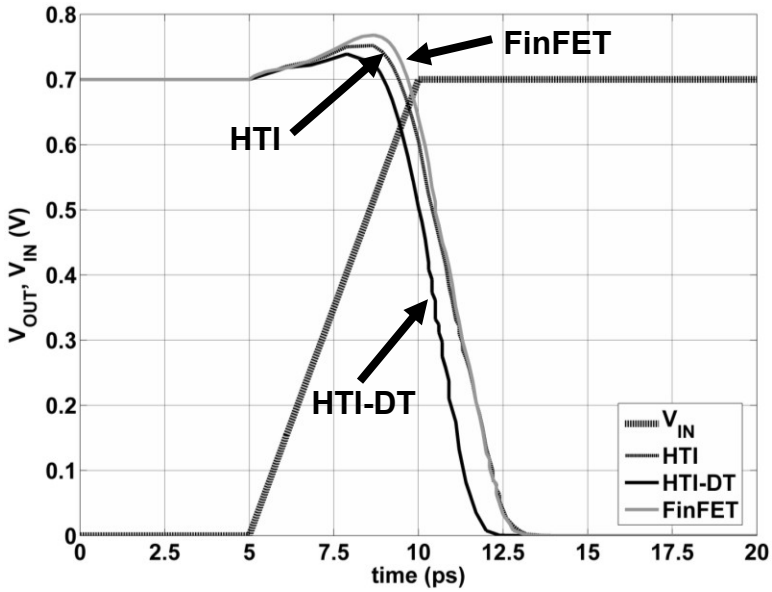

Fig. 5.12. Inverter delay simulation results comparing the HTI, HTI-DT, and FinFET inverters, all with $C_L = 2*C_{GG}$ and an air-gap low-k outer spacer. Only the pull-down operation is shown here for clarity, since the pull-up operation is symmetric. L_G, L_{SDE}, etc. are the same as for Fig. 5.11.

In principle, the Miller effect in all three structures can be reduced by replacing the SiO_2 outer spacer with something having a lower dielectric constant, as was shown in Chapter 3. To this end, an extreme example is simulated in Fig. 5.12, whereby the low-k outer spacer is air ($\varepsilon = 1$). As Fig. 5.12 shows, the performance of the HTI and HTI-DT MOSFETs improves by more than the FinFET to the point where, with an air-gap outer spacer, the FinFET is the slowest device (τ_p = 2.99 ps, 3.55 ps, 3.61 ps for HTI-DT, HTI, and FinFET, respectively). Thus, the relative

performance of the FinFET and HTI MOSFETs depends on the dielectric constant of the low-k outer spacer, being more advantageous toward the HTI MOSFETs as this outer spacer's dielectric constant is reduced.

Another consideration worth mentioning here is the gate workfunction. Throughout this study, it was assumed that any arbitrary gate workfunction could be engineered to achieve I_{OFF} at $V_{DS} = V_{DD}$ and $V_{GS} = 0$ V. For the L_{SDE}-optimized HTI NMOSFETs (Fig. 5.9), the gate workfunction is in the range of 4.63 eV to 4.66 eV (depending on L_G) and 4.51 eV to 4.53 eV for the HTI-DT NMOSFETs. For the optimized FinFETs (NMOS), the optimal gate workfunction is slightly lower, ranging from 4.67 eV to 4.78 eV. These values are all mid-gap or near mid-gap, suggesting a requirement for metal gate electrodes. Gate workfunction tunability over this range is well-demonstrated for metal/SiO_2 gate stacks [19], [20], although the small dimensions studied here may lead to integration challenges due to metal grain size limitations. This is especially a concern for FinFETs, which have a high aspect ratio active region that may lead to gate workfunction variation along the height of the fin.

5.5 Summary

In this work, an alternative MOSFET structure to the FinFET, the HTI MOSFET, has been proposed as a means of ultimate CMOS scaling without the process complexities imposed by FinFET fabrication. This structure is essentially the same as a conventional bulk MOSFET, with the exception that the trench isolation material is a high-k dielectric. This amplifies gate fringing field coupling to the active sidewalls, permitting a recess in the retrograde well doping to reduce BTBT leakage for LSTP design. The initial TCAD investigation presented here suggests that this HTI MOSFET structure, in its ideal form, is competitive with FinFETs all the way to the end of the CMOS roadmap, even for LSTP design where device subthreshold behavior is paramount.

5.6 References

[1] B. Yu, L. Chang, S. Ahmed, H. Wang, S. Bell. C.-Y. Yang, C. Tabery, C. Ho, Q. Xiang, T.-J. King, J. Bokor, C. Hu, M.-R. Lin, D. Kyser, "FinFET Scaling to 10 nm Gate Length," *IEDM Tech Dig.*, pp. 251-254, 2002.

[2] B. Doyle, B. Boyanov, S. Datta, M. Doczy, S. Hareland, B. Jin, J. Kavalieros, T. Linton, R. Rios, R. Chau, "Tri-Gate Fully-Depleted CMOS Transistors: Fabrication, Design and Layout," *VLSI Tech. Dig*, pp. 133-134, 2003.

[3] F.-L. Yang, H.-Y. Chen, F.-C. Chen, C.-C. Huang, C.-Y. Chang, H.-K. Chiu, C.-C. Lee, C.-C. Chen, H.-T. Huang, C.-J. Chen, H.-J. Tao, Y.-C. Yeo, M.-S. Liang, C. Hu, "25 nm CMOS Omega FETs," *IEDM Tech. Dig.*, pp. 255-258, 2002.

[4] J. P. Colinge, M. H. Gao, A. Romano-Rodriguez, H. Maes, C. Claeys, "Silicon-on-Insulator "Gate-All-Around Device"," *IEDM Tech. Dig.*, pp. 595-598, 1990.

[5] X. Sun, Q. Lu, V. Moroz, H. Takeuchi, G. Gebara, J. Wetzel, S. Ikeda, C. Shin, T.-J. King Liu, "Tri-Gate Bulk MOSFET Design for CMOS Scaling to the End of the Roadmap," *IEEE Elec. Dev. Lett.*, vol. 29, no. 5, pp. 491-493, May 2008.

[6] T. Iizuka, K. Y. Chiu, J. L. Moll, “”Double Threshold MOSFETs in Bird’s-Beak Free Structures,” *IEDM Tech. Dig.*, pp. 380-383, 1981.

[7] *User’s Manual for Sentaurus Device*, Synopsys Co., Mountainview, CA.

[8] X. Sun, Qiang Lu, H. Takeuchi, S. Balasubramanian, T.-J. King Liu, “Selective Enhancement of SiO_2 Etch Rate by Ar-Ion Implantation for Improved Etch Depth Control,” *Electrochem. Solid State Lett.*, vol. 10, no. 9, pp. D89-D91, 2007.

[9] M. Kito, R. Katsumata, M. Kondo, S. Ito, K. Miyano, M. Kido, H. Yasutake, Y. Nagata, N. Aoki, H. Aochi, A. Nitayama, “Vertex channel array transistor (VCAT) featuring sub-60nm high performance and highly manufacturable trench capacitor DRAM,” *VLSI Tech. Dig.*, pp. 32-33, 2005.

[10] *International Technology Roadmap for Semiconductors (ITRS)*. [Online]. Available: http://public.itrs.net.

[11] R. A. Vega, T.-J. King Liu, “A comparative study of dopant-segregated Schottky and raised source/drain double-gate MOSFETs,” *IEEE Trans. Elec. Dev.*, vol. 55, no. 10, pp. 2665-2677, Oct. 2008.

[12] S. Migita, Y. Watanabe, H. Ota, H. Ito, Y. Kamimuta, T. Nabatame, A. Toriumi, “Design and Demonstration of Very High-k (k~50) HfO_2 for Ultra-Scaled Si CMOS, *VLSI Tech. Dig.*, pp. 152-153, 2008.

[13] R. A. Vega, T.-J. King Liu, “Three-Dimensional FinFET Source/Drain and Contact Design Optimization Study,” *IEEE Trans. Elec. Dev.*, vol. 56, no. 7, pp. 1483-1492, July 2009.

[14] C. Shin, X. Sun, T.-J. King Liu, “Study of Random-Dopant-Fluctuation (RDF) Effects for the Trigate Bulk MOSFET,” *IEEE Trans. Elec. Dev.*, vol. 56, no. 7, pp. 1538-1542, July 2009.

[15] R. A. Vega, K. Liu, T.-J. King Liu, “Dopant-Segregated Schottky Source/Drain Double-Gate MOSFET Design in the Direct Source-to-Drain Tunneling Regime,” *IEEE Trans. Elec. Dev.*, vol. 56, no. 9, pp. 2016-2026, Sept. 2009.

[16] Q. Chen, L. Wang, J. D. Meindl, “Fringe-induced barrier lowering (FIBL) induced threshold voltage model for double-gate MOSFETs,” *Solid State Electron.*, vol. 49, no. 2, pp. 271-274, Feb. 2005.

[17] C.-H. Ge, C.-C. Lin, C.-H. Ko, C.-C. Huang, Y.-C. Huang, B.-W. Chan, B.-C. Perng, C.-C. Sheu, P.-Y. Tsai, L.-G. Yao, C.-L. Wu, T.-L. Lee, C.-J. Chen, C.-T. Wang, S.-C. Lin, Y.-C. Yeo, C. Hu, “Process-strained Si (PSS) CMOS technology featuring 3D strain engineering,” *IEDM Tech. Dig.*, pp. 73-76, 2003.

[18] M. Nihei, M. Horibe, A. Kawabata, Y. Awano, “Carbon nanotube vias for future LSI interconnects,” *Proceedings of the Symposium on Semiconductor Integrated Circuits and Technology*, pp. 251-253, 2004.

[19] 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, pp. 39-46, Jan. 2005.

[20] A. E.-J. Lim, J. Hou, D.-L. Kwong, Y.-C. Yeo, “Manipulating Interface Dipoles of Opposing Polarity for Work Function Engineering within a Single Metal Gate Stack,” *IEDM Tech. Dig.*, pp. 33-36, 2008.

Chapter 6

Implant-to-Silicide Process Technology

6.1 Introduction

DSS MOSFETs have been experimentally demonstrated using a variety of substrates and geometries [1]-[11], including planar bulk [2], [5], [7], planar SOI [1], [4], [8]-[11], and SOI MuGFET architectures [3], [6], [10], [12]. However, no experimental work to-date has taken place to understand design optimization paths for DSS MOSFETs. Instead, most published works on the topic have demonstrated variations of source/drain silicide material [1], [2], [6], silicide capping layers [13], and DSS doping schemes such as dopant pile-up (also known as implant before silicide (IBS) [12], [14] or silicidation-induced dopant segregation (SIDS) [15]) or implant-to-silicide (ITS [12], also known as implant after silicide (IAS) [14] or silicide as diffusion source (SADS) [15]). In addition to the usual design optimization parameters which include sidewall spacer thickness, silicide thickness, etc., the SDE junction depth ($X_{j,SDE}$) is a critical component of DSS MOSFET optimization at aggressive scales [16]-[18], and so it must be understood if/how $X_{j,SDE}$ can be modulated for DSS MOSFETs in a process environment.

For the ITS method, there are three processes that must take place in order to form a DSS junction: 1) dopants must diffuse within the silicide toward the silicide-silicon interface -- dopant diffusivity in silicides has been shown to be very high due to grain boundary diffusion [19], [20]; 2) dopants must segregate out of the silicide -- this segregation effect has been shown to be driven by the interfacial dipole creating energetically favorable substitutional bonding sites for dopants at the silicon side of the silicide-silicon interface [14]; and 3) dopants must diffuse some distance into the Si. It is this third process, the diffusion process, which is not well understood in the formation of DSS junctions. For example, in [15], it was shown that NiSi and PtSi provide for different dopant profiles when ITS is used to form DSS junctions, but no explanation was offered. The presence of excess vacancies or other point defects, somehow linked to the presence of silicide, has been proposed as the underlying mechanism [21], [22]. This has not been validated experimentally, neither has there been any reported work to determine whether there is any way to achieve SDE profiles with NiSi as sharp as those achieved with PtSi [15]. This would be beneficial, because NiSi is much easier and less expensive to process than PtSi (due to the high etch selectivity between NiSi and unreacted Ni).

In this chapter, experimental results are presented to provide a clear understanding of the factors that determine the depth of doped junctions formed by ITS, to facilitate DSS MOSFET process optimization. DSS diodes and MOSFETs are fabricated with ITS and shown to have tunable SDE regions by fluorine pre-silicidation ion implant (F-PSII).

6.2 DSS Diode Fabrication

The starting substrates were lightly-doped (~$1x10^{15}$ cm^{-3}) n-type and p-type wafers. 100 nm of thermal oxide was grown by dry oxidation and then patterned to form 1 cm^2 square holes using a 5:1 buffered HF wet etch. Some samples then underwent a fluorine pre-silicidation ion implant (F-PSII) with dose, energy, and tilt angle of $1x10^{15}$ cm^{-2}, 20 keV, and 0 °, respectively. Subsequently, an *in-situ* sputter pre-clean was performed followed by sputter deposition of 30 nm of Ni. The wafers were then annealed in an oven at 300 °C for 5 min in N_2, to form Ni_2Si. After the unreacted Ni was selectively removed in a heated H_2SO_4+H_2O_2 solution, NiSi was formed by rapid thermal annealing (RTA) at 500 °C for 1 minute. ITS was then performed at $1x10^{15}$ cm^{-2} dose and 0 ° tilt, using either As (35 keV), P (20 keV), or BF_2 (25 keV) as the implant species. The implant energies were selected based on SRIM simulation [23] to achieve similar as-implanted profiles (Fig. 6.1). Each wafer was then cleaved into four quarter pieces, each containing 10-20 die; each die is a 1 cm^2 diode and the diodes with the lowest reverse bias leakage were chosen for capacitance measurements and secondary ion mass spectrometry (SIMS) analysis. These quarter pieces were then placed onto a pocket wafer and either subjected to RTA in N_2 at 500, 600, or 700 °C for 1 min, or extended RTA annealing at 600 °C for 30 min.

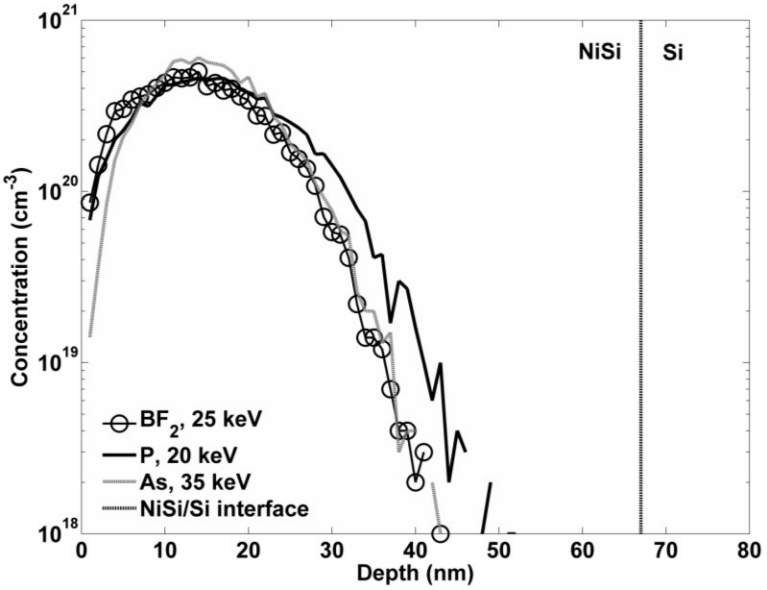

Fig. 6.1. SRIM simulation results for ITS into NiSi, with a 10,000 ion count and full damage cascades, at $1x10^{15}$ dose for BF_2 (25 keV), P (20 keV) and As (35 keV). The BF_2 simulation is actually B at 5.61 keV, since BF_2 implants cannot be simulated with SRIM.

6.3 DSS SDE Formation Using ITS

As mentioned previously, it was reported in [15] that DSS junctions formed by ITS have sharper profiles when PtSi is used rather than NiSi. It is plausible that this is related to the difference in thermal stability of PtSi *vs.* NiSi, the former being more stable at higher temperatures [24]. Indeed, it was shown in [25] that Ni atoms are rejected from the silicide and diffuse into the Si substrate when NiSi is annealed for extended periods of time at sub-

agglomeration temperatures. This process may represent the onset of agglomeration, as NiSi grains with high interface energy [26] (either between NiSi grains or between a single NiSi grain and the Si) evolve toward a lower energy state through a combination of secondary grain growth and atomic Ni rejection from the silicide. The SIMS data in [14] show that the Ni profiles have almost the same abruptness as the dopant profiles, which lends support to the theory that silicide thermal instability and correlated point defect distribution in Si affect the dopant profile. In short, a post-ITS anneal can result in some metal atom rejection from the silicide and diffusion into the Si, which mediates dopant diffusion and the formation of DSS junctions. If the metal atoms diffuse substitutionally, the local Si bonding energy will be reduced and the activation energy for dopant diffusion will drop. An excess of vacancies may also occur if substitutional metal diffusion is faster than the rate of pairing between vacancies and silicon self-interstitials. If the metal atoms diffuse interstitially, their relatively large size may generate enough stress within the Si lattice to break Si-Si bonds and give rise to vacancies, again enhancing dopant diffusion, especially if the metal atoms form clusters, as reported in [25].

It would naturally follow that, in order to "tune" the SDE junction depth in DSS structures for a given process thermal budget, one must control the thermal stability of the silicide. This is most simplistically achieved by changing the silicide material, as in [15]; however, this provides for a very limited range of tunability due to limited material choices. A variation on this method would be silicide alloying. For example, NiPtSi, with low-to-moderate Pt content, has been a popular material choice lately [27]-[29] because it has the process integration advantages of NiSi and the thermal stability advantages of PtSi. In principle, varying the Pt content should result in tuning of the DSS junction depth; however, this approach is limited by the fact that NiPtSi is comprised of NiSi and PtSi grains [28], [29]. This gives rise to variations larger than those which ordinarily result from a distribution of grain orientations within a single silicide, as now the grain orientation variations of the two silicides are combined. Also, for ultra-thin or narrow transistor structures such as fully depleted silicon-on-insulator (FD-SOI) MOSFETs or FinFETs, which can have body thickness comparable to the average grain size in the silicide, the DSS tunability becomes a probability that either a NiSi or PtSi grain will abut the channel region and determine the SDE junction depth. Thus, a method of modulating the thermal stability of a single silicide material for tuning DSS junctions with minimal variability is preferable.

F co-implantation (with dopants) into silicide has been demonstrated to improve the thermal stability of NiSi [30]; however, for ITS processing, this process results in a race condition between F segregation at the NiSi/Si interface and Ni diffusion from the same interface and therefore reduced DSS tunability. An alternative approach, F-PSII, was reported in [31], [32] to exert considerable control over the Ni diffusion profile during thermal treatment, due to NiSi/Si interface dangling bond passivation and a concomitant reduction in interface energy. The same study reported similar results with N (N-PSII), due instead to grain boundary passivation, although the effect was not as strong as that of F-PSII. F-PSII was chosen for the present study because of its efficacy in controlling Ni diffusion, and also because it does not affect the silicide phase, in contrast to N-PSII which tends to promote the formation of $NiSi_2$ at the NiSi/Si interface [33].

Fig. 6.2 shows SIMS depth profiles for (a) BF_2, (b) P, and (c) As implanted samples with a 600 °C/30 min post-ITS anneal. (A 30 min anneal was chosen to exacerbate any difference in junction depth between the splits with and without F-PSII.) Since the Ni depth profiles are broad, a depth correction factor could not be applied to the NiSi region, and so Fig. 6.2 shows the dopant segregated peaks at a shallower depth (~ 25 nm) than they actually are (~ 67 nm).

Nevertheless, Fig. 6.2 shows very clearly that the dopant profiles are contained entirely within the region of the Ni profiles. In particular for Figs. 6.2 (a) and (b), as B and P have a higher diffusivity than As under the same annealing conditions, F-PSII suppresses the junction profile far from the NiSi/Si interface, exactly as predicted. At the same time, for Figs. 6.2 (a) – (c), F-PSII enhances dopant diffusion close to the NiSi/Si interface. This local enhancement is due to implant damage from the F-PSII, as F tends to enhance dopant diffusion in sub-amorphous Si [34]. Thus, with F-PSII designed so as to minimize F implant damage at the NiSi/Si interface (*i.e.*, lower implant energy and/or thicker Ni), and/or a damage-less process such as N co-plasma [35] during Ni sputtering, the SDE junctions formed by ITS can be sharpened due to the tighter spatial distribution of Ni.

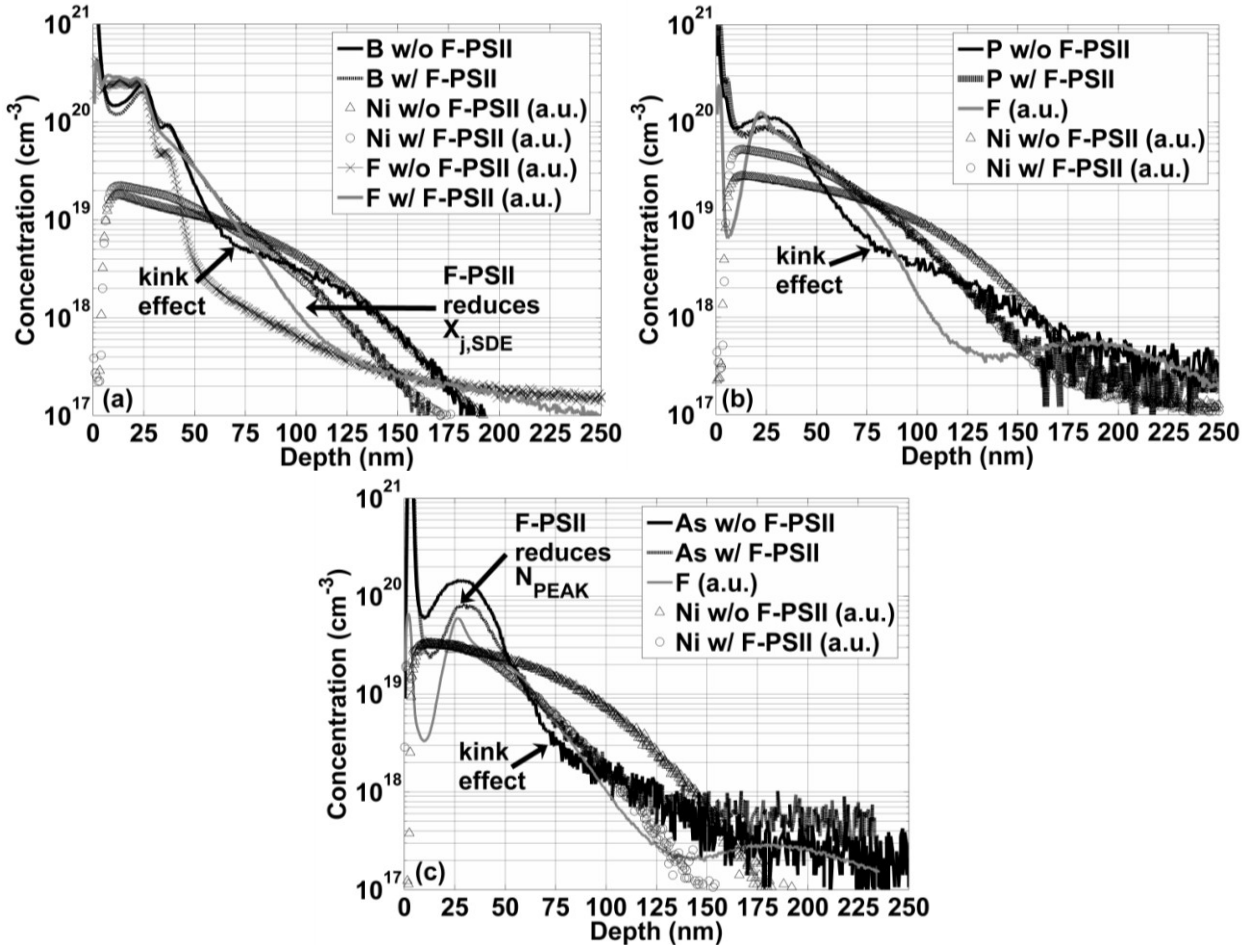

Fig. 6.2. SIMS depth profiles for (a) BF_2, (b) P, and (c) As-implanted samples, with and without F-PSII. All samples were annealed at 600 °C for 30 min.

Two other features are apparent in Fig. 6.2. The first is that F-PSII reduces the segregated dopant concentration at the NiSi/Si interface (Table I below). This is because both F and the segregated dopant atoms compete for substitutional bonding sites at the NiSi/Si interface. With F already segregated at the interface before the dopants are implanted into NiSi, the post-anneal interfacial dopant concentration drops. One would argue that this may impose a trade-off between junction abruptness and SBH (and therefore specific contact resistivity ρ_c, which drops as the dopant concentration increases [36], [37]); however, this is not necessarily the case, and is

covered in more detail in the next subsection.

The second feature is the kink seen in each of the dopant profiles, at a depth of ~ 75 nm, for the samples without F-PSII (which is less evident in the samples with F-PSII, due to the F implant damage, but otherwise would still be expected). This indicates that there are two diffusion regions during the post-ITS anneal. The first region, close to the NiSi/Si interface, has excess vacancies due to atomic Ni diffusion into the substrate and is henceforth referred to as the atomic Ni region. The second region, further from the NiSi/Si interface, has excess vacancies due to Ni clustering, as in [25], and is henceforth referred to as the clustered Ni region. (The distance from the NiSi/Si interface at which this clustering takes place is the cluster length L_c.) The Ni clusters are much larger in size than atomic Ni and so the excess vacancy concentration in this region is much higher than in the atomic Ni region. Thus, as the segregated dopants diffuse away from the NiSi/Si interface, their diffusivity is relatively low (in the atomic Ni region) and so the junction profile will be fairly abrupt (*e.g.*, 20.43 nm/dec in Fig. 6.2(a) for the B profile without F-PSII, at ~ 50 nm depth). As the dopants diffuse a distance beyond L_c, their diffusivity increases considerably (in the clustered Ni region), resulting in a broadening of the dopant profile (*e.g.*, 50.38 nm/dec in Fig. 6.2(a) for the B profile without F-PSII, at ~ 175 nm depth).

An exact value for L_c may be difficult to define, since Ni clusters may exist in various sizes, leading to a transition region between the atomic Ni region and the clustered Ni region. However, the distinct kink in the dopant profiles suggests that L_c may be reasonably defined as the distance from the NiSi/Si interface (*i.e.*, the dopant profile peak) to the kink in the dopant profile. In the example of Fig. 2(a) for BF_2 without F-PSII, the B peak is at 25 nm while the kink occurs at ~65 nm, meaning L_c ~ 40 nm. The reason why this effect has not been observed in previously published SIMS data on DSS junctions formed by ITS [14], [15] could be that Ni clustering increases with time. In other words, L_c is constant, but the Ni profile spreads out, such that the Ni concentration and therefore Ni cluster concentration at L_c increases as the anneal time increases. The post-ITS anneal time in [14] was not reported, but it was only 30 sec in [15]; in comparison, the data in Fig. 6.2 are for a 30 min anneal. Thus, containing the dopant profile within L_c or reducing the Ni cluster concentration at L_c (both of which require short anneals) should result in a smooth dopant profile. This may be extended further with F-PSII, which is expected to increase L_c due to F gettering of stray Ni atoms [31], [32].

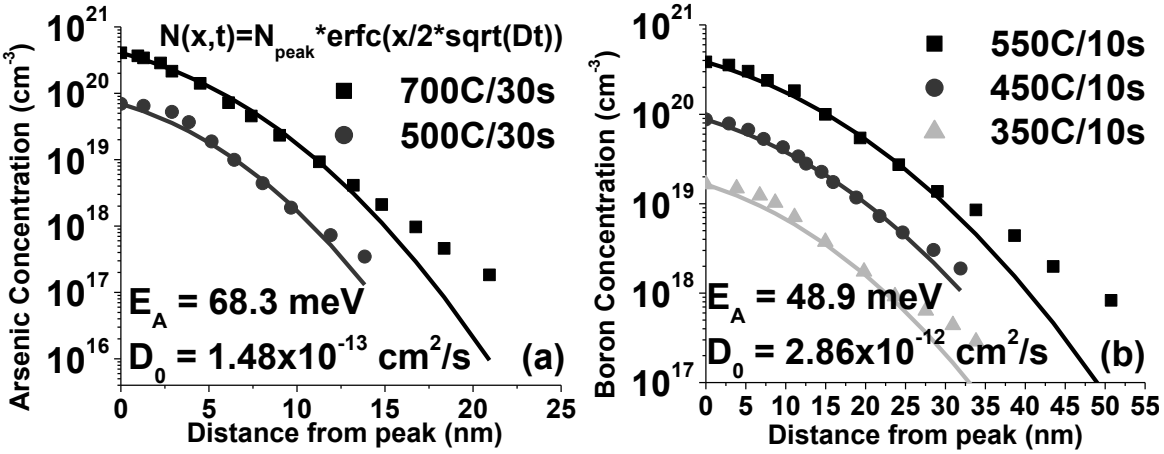

Fig. 6.3. Modeled erfc profiles (solid lines) vs. data for (a) As/PtSi (Ref. [7]) and (b) B/NiSi (Ref. [9]), for ITS processing.

For short anneal times such that $X_{j,SDE} < L_c$, the SDE profile should follow erfc (solid-source diffusion) behavior, since the dopants are effectively "dragged" into the silicon by the diffusing Ni, which itself comes from a semi-infinite source of Ni atoms in the NiSi layer. This is shown in Fig. 6.3, where erfc model curves are plotted against previously published data (the data points in Fig. 6.3 were traced by hand). The experimental data at multiple temperatures permits extraction of the activation energy E_A and upper diffusivity limit D_0 for the presented dopant/silicide combinations in Fig. 6.3. For both cases (*i.e.*, sharp As/PtSi junctions and broad B/NiSi junctions), E_A is very low (50-70 meV), while D_0 shows the expected difference between B and As (~20x).

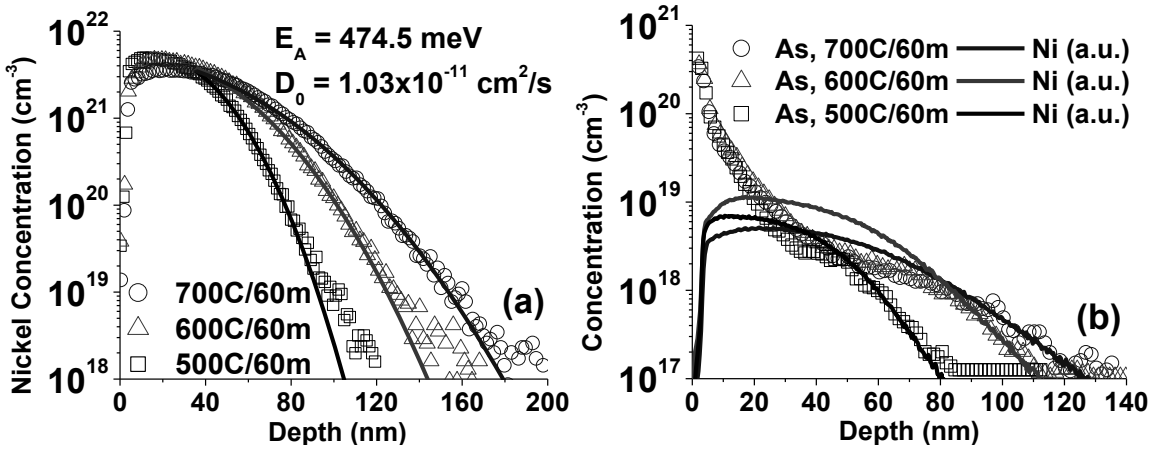

Fig. 6.4. (a) modeled erfc profiles (solid lines) vs. Ni SIMS for long anneal times and (b) overlay of Ni and As SIMS for the same anneals, showing the kink effect in the As profile due to Ni clustering. For these samples, 16 nm of Ni was deposited, as opposed to 30 nm in Fig. 6.2, and the As was implanted at 10 keV, 0 º tilt, and $1x10^{15}$ cm^{-2} dose.

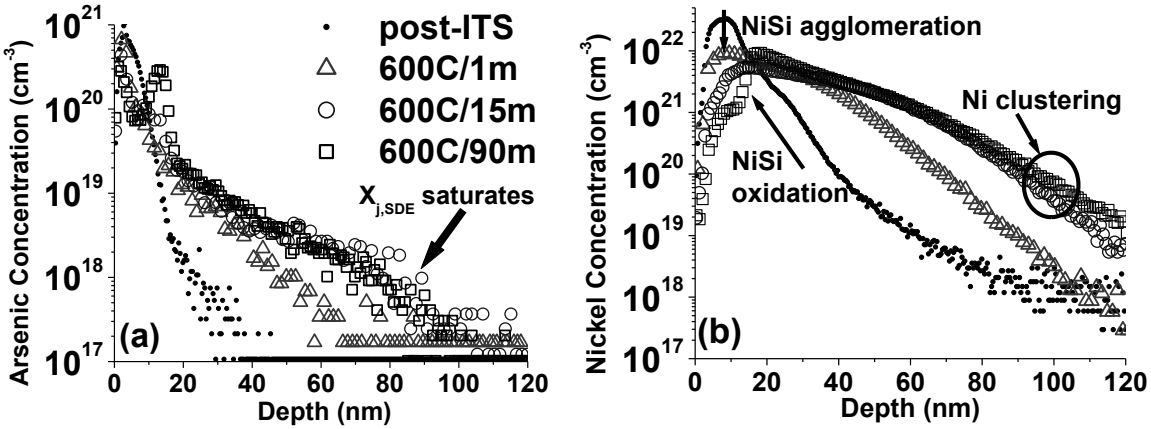

Fig. 6.5. (a) As and (b) Ni SIMS after ITS and anneal for various times. The sharp As peak at ~15-20 nm is due to inadvertent NiSi oxidation during the RTA cool-down phase. For these samples, 16 nm of Ni was deposited, as opposed to 30 nm in Fig. 6.2, and the As was implanted at 10 keV, 0 º tilt, and $1x10^{15}$ cm^{-2} dose.

It would therefore seem that E_A is associated with the process of metal atom rejection from the silicide at sub-agglomeration temperatures. This is because, for long anneal times whereby clustered Ni diffusion dominates, E_A for Ni diffusion is predictably higher, at 474.5 meV (Fig. 6.4(a)). Regardless of the anneal temperature, the dopant profile follows the Ni profile (Fig. 6.4(b)), and $X_{j,SDE}$ saturates for long enough anneal times. This saturation in $X_{j,SDE}$ is shown

in Fig. 6.5(a) and is due to Ni clustering which increases E_A and retards Ni diffusion (Fig. 6.5(b)). It is noteworthy that the samples prepared for Figs. 6.4 and 6.5 had 16 nm of Ni sputter deposited rather than the 30 nm used in Fig. 6.2. This resulted in a thinner NiSi layer (~ 36 nm), which readily agglomerated (Fig. 6.6(b)). What this means is that dopant diffusion into the substrate is independent of the morphological state of the silicide layer, whose only function in this regard is to act as an initial source of metal atoms.

Assuming that D_0 is not affected by fluorine, the change in E_A due to F-PSII can be extracted using

$$\frac{X_j}{X_{j,F-PSII}} = \sqrt{\frac{D}{D_{F=PSII}}} \Longrightarrow \Delta E_A = kT * \ln\left[\left(\frac{X_j}{X_{j,F-PSII}}\right)^2\right] \tag{6.1}$$

where X_j and D are respectively the junction depth and dopant diffusivity without F-PSII, and $X_{j,F\text{-}PSII}$ and $D_{F\text{-}PSII}$ are respectively the junction depth and dopant diffusivity with F-PSII. From the ratio of junction depths in Fig. 6.2 (taken at a dopant concentration of $1x10^{17}$-$1x10^{18}$ cm^{-3}, or equivalently from the Ni profiles in Fig. 6.2), E_A for B and P diffusion (From Fig. 6.2) in Si increase with F-PSII, by 21.5 meV and 23.8 meV, respectively. The increase in E_A (ΔE_A) for Ni diffusion is smaller for BF_2-implanted NiSi (26.2 meV) than for P-implanted NiSi (35.6 meV) because of the extra F introduced by the BF_2 implant. For the As-implanted sample in Fig. 6.2(c), there is not enough of a change in the dopant profile with F-PSII to measure ΔE_A, which is anomalous if one considers the results for As-implanted samples in Figs. 6.3-6.5 (the cause for this remains unclear, but may be due to thermal isolation via the pocket wafer during RTA). However, ΔE_A due to F-PSII for the Ni profiles in Fig. 6.2(c) is found to be 43.7 meV, which is considerably higher than ΔE_A for the BF_2- and P-implanted samples. A likely explanation for this is that As-doped NiSi is less thermally stable than B-doped or P-doped NiSi [38]-[40], meaning F-PSII should result in the largest ΔE_A for the As-implanted sample.

These findings support the theory that metal atom rejection from the silicide is the dominant factor in determining the doping profiles. They also suggest that NiSi ITS can be competitive with PtSi ITS since, *e.g.*, F-PSII increases E_A to ~70.4 meV for BF_2/NiSi ITS and short anneals.

6.4 Diode Capacitance-Voltage Analysis

The minority carrier barrier height ϕ_B was extracted using the $1/C^2$ vs. V methodology outlined in [41]. Since it has now been established that DSS junction formation is a direct function of the spatial Ni (and therefore trap) distribution, it would follow that ϕ_B extraction by C-V cannot be performed at any single frequency. The presence of these traps will increase the minority carrier response time, thus reducing the measured capacitance at high frequency. This results in an extracted built-in voltage (V_{bi}) and therefore ϕ_B that is artificially high, an example of which is shown in Fig. 6.6. Thus, the measurement frequency must be swept over a wide range to find the appropriate frequency at which the minority carrier response time is negligible and the correct ϕ_B value can be extracted. The resulting ϕ_B vs. frequency curves are shown in Fig. 6.7.

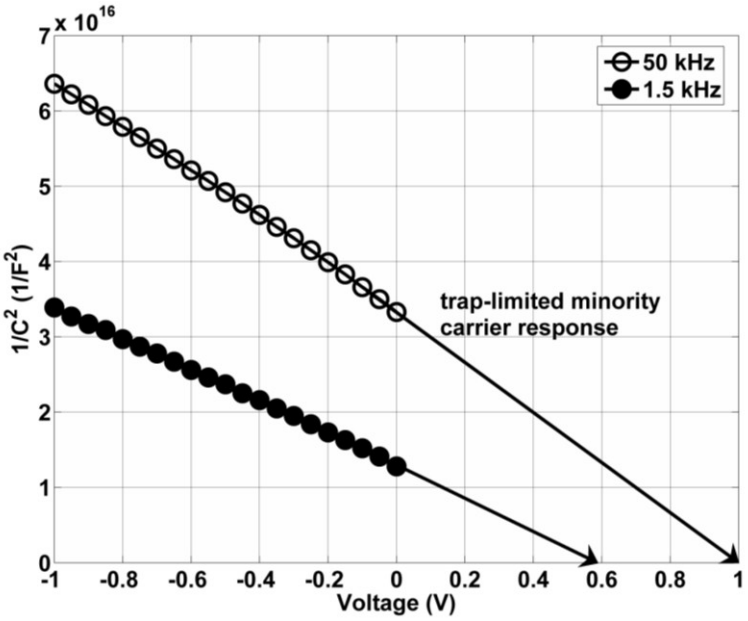

Fig. 6.6. Example plot of $1/C^2$ vs. V for ITS with BF_2 and without F-PSII, at 50 kHz and 1.5 kHz measurement frequency.

Several interesting features emerge when observing the effect of F-PSII, anneal time, and doping polarity on the ϕ_B vs. frequency curves in Fig. 6.7. In Fig. 6.7(a), for the curves without F-PSII, increasing the anneal time slightly increases the minority carrier ϕ_B from 0.9 eV at 2.5 kHz to 0.94 eV at 1.5 kHz. This suggests that increasing the post-ITS anneal time will result in a higher B concentration at the NiSi/Si interface and therefore higher electron ϕ_B. This is only observed for B and not P or As, and so is likely due to F from the BF_2 implant contained within the NiSi layer (Fig. 6.2(a)) reducing B diffusivity within the NiSi, thereby increasing the time required for the interfacial B concentration to saturate during the post-ITS anneal. A meaningful prediction from this would be that ITS with B^{11} rather than BF_2 would not show this effect, thus permitting both DSS NMOS and PMOS devices to have the same optimal thermal budget for the post-ITS anneal. (Individual tuning of the NMOS and PMOS SDE regions can be achieved by separate F-PSII.)

What is seen from all of the curves in Fig. 6.7, but in particular in Fig. 6.7(a), is that longer anneal times shift the ϕ_B vs. frequency curve to lower frequencies. This is indicative of the increase in spatial Ni (and therefore trap) distribution shown in Fig. 6.2. One could infer that the more the ϕ_B vs. frequency curve is shifted to lower frequency, the deeper the SDE junction. However, this is not always the case. Among the curves for 1 min anneal time, the curve with F-PSII is shifted to a lower frequency than the curve without F-PSII. This is a consequence of the implant damage from F-PSII. However, for 30 min anneal time, the deepest traps in the Si are those formed by Ni diffusion, and so the downward frequency shift is truly indicative of a deeper SDE junction.

The frequency shift is more evident in Fig. 6.7(a) than in Figs. 6.7(b) and (c), because the extracted ϕ_B begins to drop at ~ 10-100 KHz for the p^+n diodes (whereas it does not begin to drop until the frequency is reduced to ~ 1 kHz for the n^+p diodes). The reason for this difference between the BF_2 and As/P samples may be due to deep trap states located closer to the valence band, for which minority carriers have a short response time for p^+n diodes but a long response time for n^+p diodes. It is possible that these deep traps are associated with vacancies formed by

Ni diffusion into the Si. They may also be the result of the Ni atoms/clusters themselves, which give rise to both acceptor-like (E_c – 0.47 eV) and donor-like (E_V + 0.18 eV) states in silicon [42]. This is more likely, considering that vacancies should be quickly filled by dopant atoms.

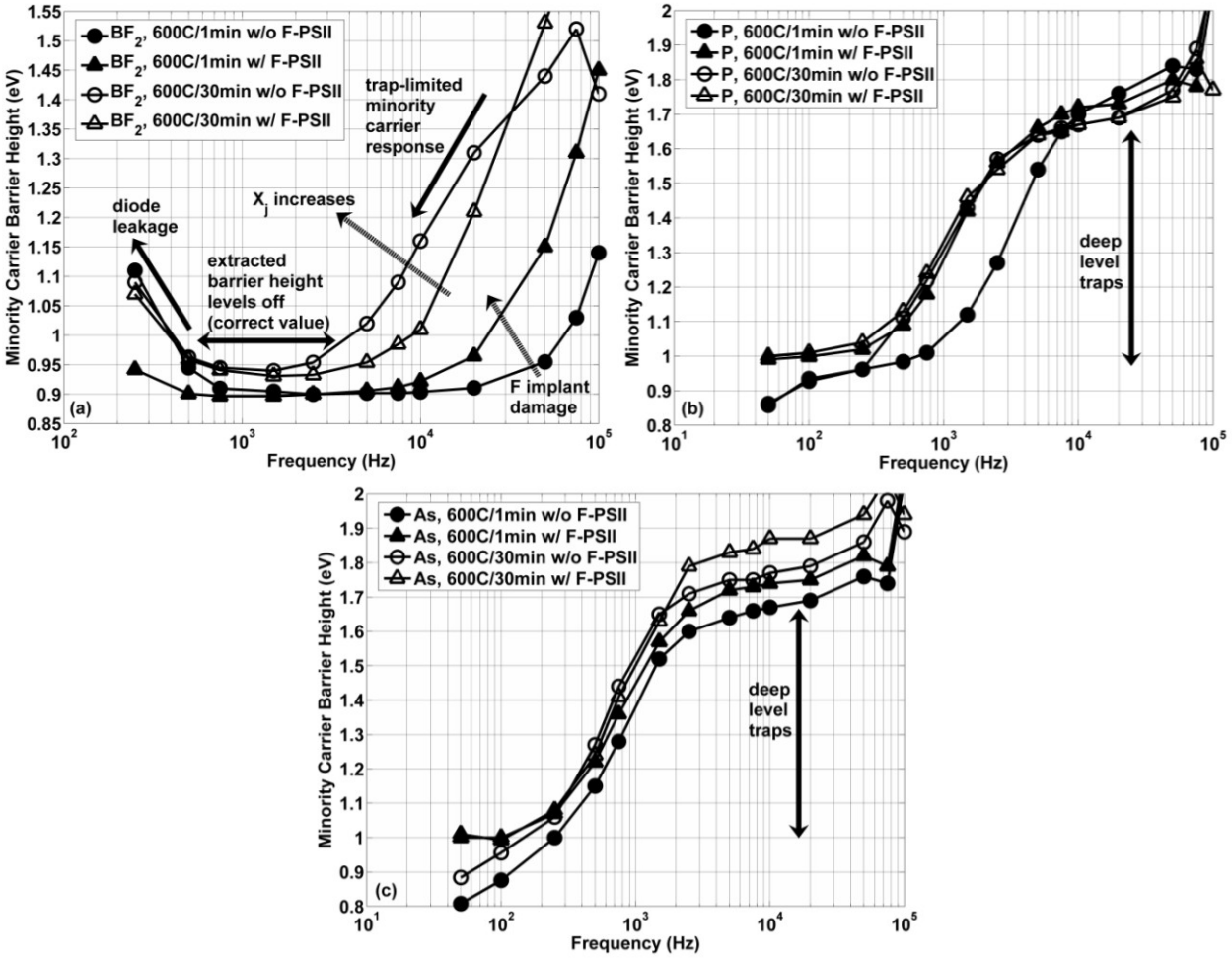

Fig. 6.7. Minority carrier ϕ_B vs. measurement frequency for (a) BF_2, (b) P, and (c) As-implanted samples, with and without F-PSII. Post-ITS anneals were performed at 600 °C for 1 min or 30 min. The data for 30 min are from the exact same die as measured for SIMS depth profiles in Fig. 6.2.

A final point to note regarding Fig. 6.7 is that F-PSII does not affect ϕ_B for the p^+n diodes but notably increases ϕ_B for the n^+p diodes (Table I). This is opposite to expectations for reduced ϕ_B in Figs. 6.7(b) and (c), based on the observed reduction in interfacial dopant concentration with F-PSII (Fig. 6.2). This suggests either that the segregated F atoms reduce the electron ϕ_B by an amount that is greater than the decrease caused by a drop in interfacial dopant concentration, or that some other effect is taking place. In [43], F passivation of NiSi/Si interfaces was found to have no significant effect on ϕ_B, regardless of implant dose, so it is clear that some other effect is the cause. One possibility is that F implant damage reduces the electron ϕ_B, similar to the effect of Ar plasma pre-amorphization on $ErSi_{1.7}$ contacts [44]; however, a corresponding reduction in ϕ_B is not measured in Fig. 6.7(a) for the p^+n diodes. Also, the ϕ_B shift in [44] is only ~ 20 meV, whereas in Figs 6.7(b) and (c) the ϕ_B shift is ~ 50-100 meV. A more likely possibility is that the

reduced spatial Ni distribution itself, due to F-PSII, contributes to an improvement in the efficiency with which n-type dopants can reduce (increase) the electron (hole) ϕ_B. This is consistent with a reduction in spatial distribution of acceptor-like states which must be screened by n-type dopants in order to reduce the electron ϕ_B.

In order to estimate the majority carrier ϕ_B, it is typical to subtract the extracted minority carrier ϕ_B from the bandgap energy E_G. However, for DSS junctions, the interface dopant concentration is very high and so using 1.12 eV as E_G for Si is inaccurate, due to bandgap narrowing (BGN) at high doping levels. To this end, BGN is calculated herein using the peak doping concentration at the DSS junction and the models in [45] for n-type and p-type Si BGN. This is the first time BGN has been included in the analysis of DSS junctions, and the results are shown in Table I. For the samples without F-PSII, the majority carrier ϕ_B is, at most, 0.105 eV. With F-PSII, all extracted values for ϕ_B are zero or near-zero. That both the n-type and p-type contacts achieve extremely low ϕ_B with ITS processing indicates that DSS MOSFETs can be designed to not have a SB contact so that they essentially are aggressively silicided conventional MOSFETs (with ohmic source/drain contacts); as a result, the parasitic resistance in these structures will be due entirely to the sheet resistances of the source/drain silicide and SDE regions. The data in Table I also explain why the minority carrier ϕ_B extracted for the p^+n diodes does not reach ~ 1 eV as it does for the n^+p diodes. The answer lies within the larger BGN that occurs in p-type Si at the same doping level, due to the different subband nature for p-type vs. n-type Si [45]. For this specific reason, a direct comparison of the minority carrier ϕ_B values for n^+p vs. p^+n DSS diodes is misleading due to the differing degrees of BGN for n-type and p-type contacts.

As Table II shows, the minority carrier ϕ_B for the n^+p diodes with F-PSII remains at ~ 1 eV, regardless of post-ITS anneal temperature, for a 1 min post-ITS anneal. (500 °C data could not be obtained for P-implanted samples, due to very high leakage.) For the p^+n diodes without F-PSII, the minority carrier ϕ_B is independent of post-ITS anneal temperature; however, with F-PSII, it drops with increasing anneal temperature. This could be due to an increase in F activation at and near the NiSi/Si interface, which would reduce the peak B concentration. This discovery reveals another advantage of F-PSII beyond its effect on reducing the spatial Ni distribution and hence the SDE junction depth in DSS MOSFETs for a given post-ITS anneal thermal budget, in that lower temperatures for the post-ITS anneal can be utilized to reduce the SDE junction depth even further, at no cost to contact resistance. This is very valuable for DSS MOSFET process optimization, not only because it can improve MOSFET scalability and performance, but also because it can reduce performance variation. In [46], it was noted that the variation in specific contact resistivity ρ_c due to random dopant fluctuation (RDF) will be very small at and near the end of the CMOS technology roadmap, if the dopant concentration at the contact interface is sufficiently large. However, it was also noted in [46] that this may lead to a requirement for co-optimization of ρ_c variation and threshold voltage V_t variation, since these have opposite dependence on SDE doping concentration. By reducing the post-ITS anneal temperature to reduce the SDE dopant concentration [14], [15] without any penalty in ϕ_B, the ρ_c and V_t co-optimization design space is increased, to allow for both reduced performance variation and increased nominal device performance.

6.5 Diode Current-Voltage Analysis

The aforementioned benefit of F-PSII (permitting a reduction in post-ITS anneal temperature without sacrificing device performance) is contingent upon the DSS diode leakage not increasing as the post-ITS anneal temperature is reduced. An increase in leakage with reduced anneal temperature was reported in [15] and generally is correlated with a reduction in minority carrier ϕ_B. For sufficiently low anneal temperatures (or more accurately, low enough thermal budgets) such that $X_{j,SDE}$ is very small and/or the SDE dopant concentration is very low, tunneling through the SDE region will also contribute to diode leakage [17], in a manner very similar to SB leakage. This may explain why the diode leakage was too high for the P-implanted samples which underwent only a 500 °C/1min post-ITS anneal, preventing C-V analysis to obtain ϕ_B information.

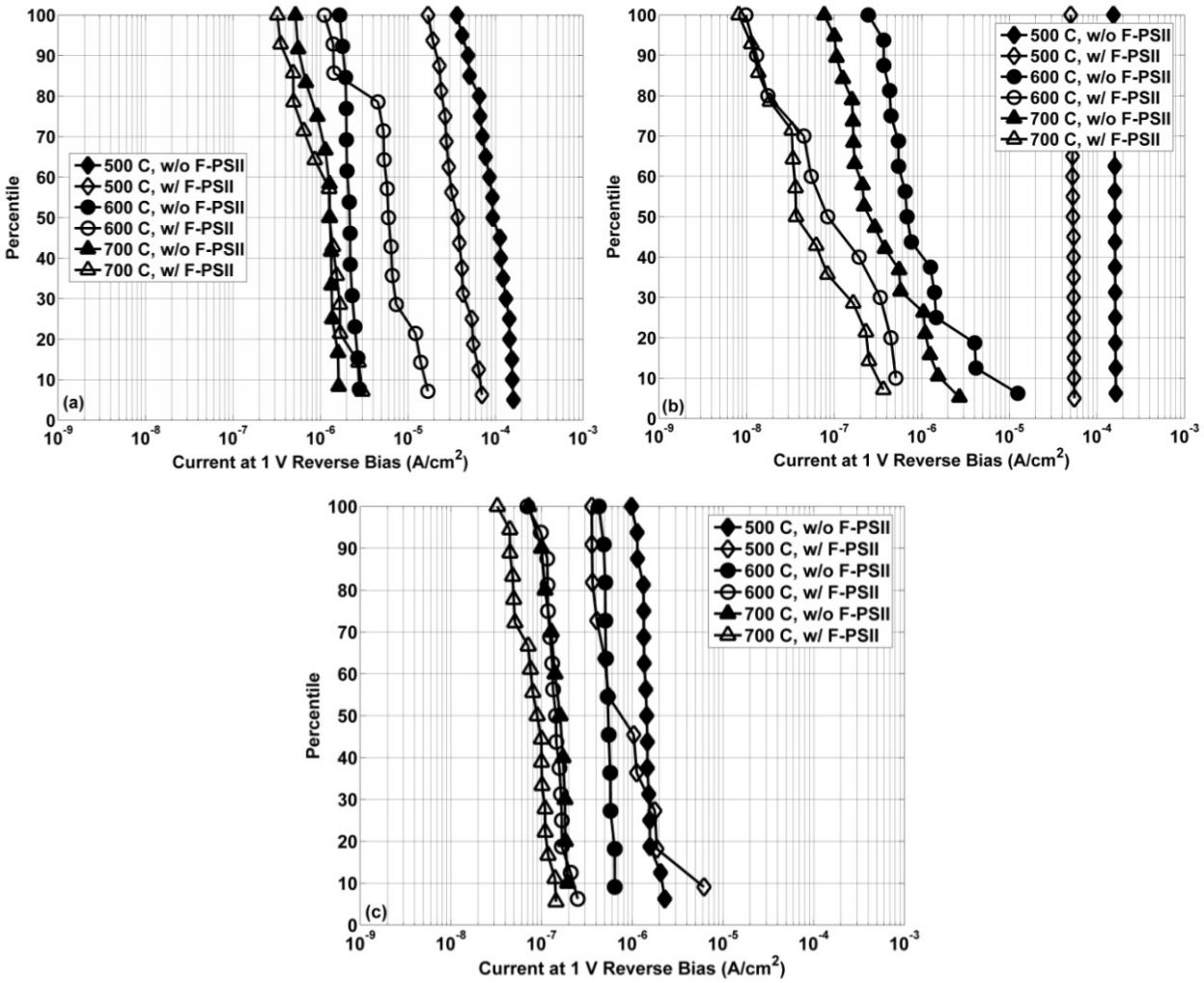

Fig. 6.8. Cumulative distribution of reverse bias diode leakage for (a) BF_2, (b) P, and (c) As-implanted samples, with or without F-PSII. All samples were annealed at 500, 600, or 700 °C for 1 min.

Fig. 6.8 shows the cumulative distributions of reverse bias diode leakage for the DSS samples annealed at 500, 600, or 700 °C for 1 min, with or without F-PSII. Leakage current for the p^+n

diodes is higher than for the n^+p diodes by ~ 1 decade, although it is noteworthy that the diode on-state current (not shown) is also higher by ~ 1 decade, so a direct I-V comparison between the p^+n and n^+p diodes would be misleading. (This may well be the result of different series resistance in the n-type and p-type wafers used to fabricate the diodes.) Nevertheless, in all cases, an increase in anneal temperature reduces the diode leakage, as expected, and is consistent with the trend reported in [15]. For the p^+n samples, there is no clear case to be made for F-PSII reducing or increasing diode leakage, perhaps due to competition between F activation and B segregation at the NiSi/Si interface. However, with F-PSII and the same post-ITS thermal budget, the diode leakage is reduced for the n^+p samples by an amount equivalent to annealing the sample without F-PSII at a temperature at least 100 °C higher.

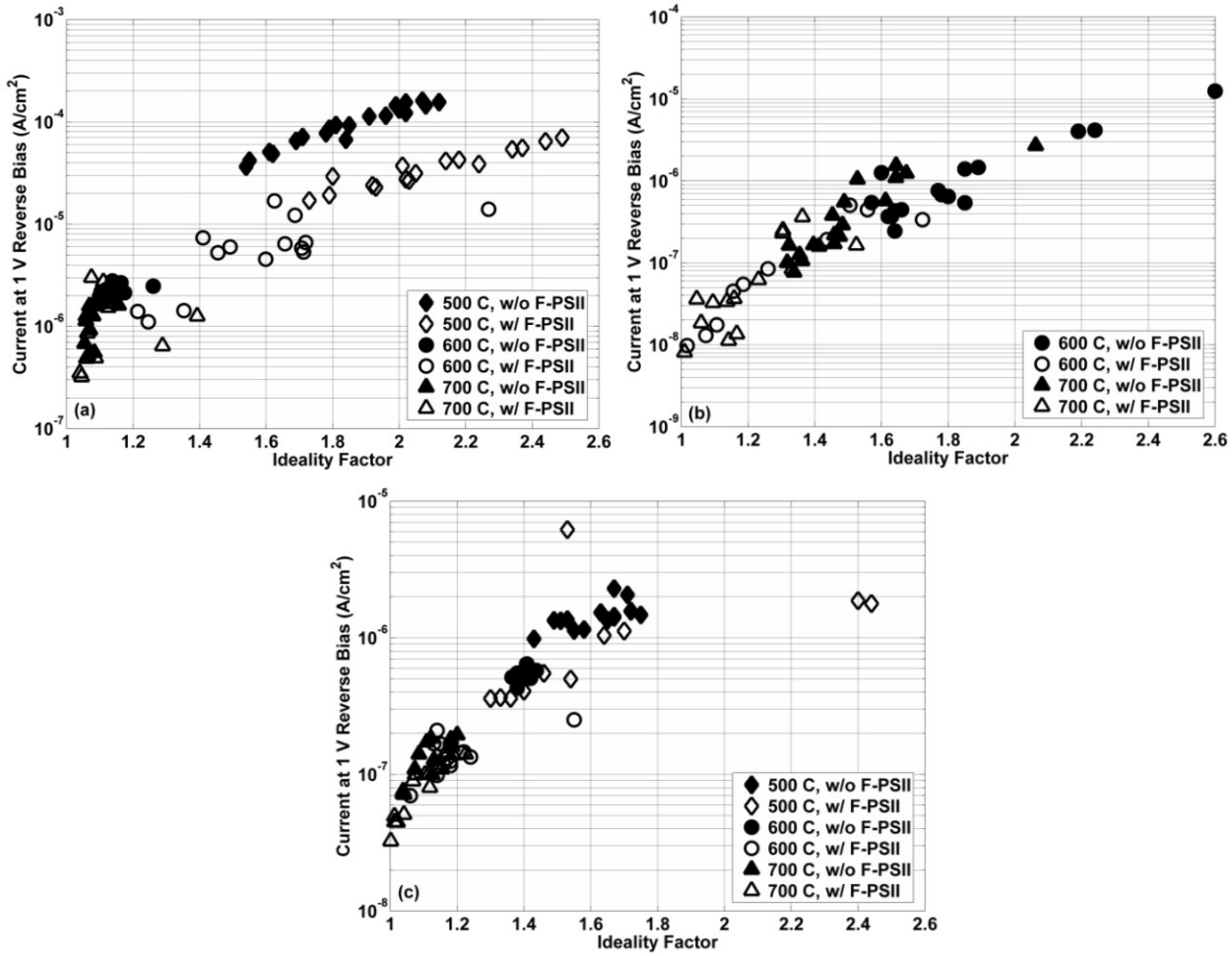

Fig. 6.9. Leakage current at 1 V reverse bias vs. diode ideality factor for (a) BF_2, (b) P, and (c) As-implanted samples, with or without F-PSII. All samples were annealed at 500, 600, or 700 °C for 1 min.

The reduction in diode leakage with F-PSII, for the n^+p diodes, correlates to the increase in minority carrier ϕ_B shown in Table II. However, it is noteworthy that in some cases, for a given process split (*e.g.*, As with F-PSII), the diode leakage drops with increasing annealing temperature despite the minority carrier ϕ_B being constant over this temperature range. Thus it would seem that something other than a change in minority carrier ϕ_B is the actual cause of leakage reduction with increasing post-ITS anneal temperature. Fig. 6.9 shows scatter plots of diode leakage at 1 V reverse bias versus the corresponding diode ideality factor. In all cases,

regardless of anneal temperature or whether F-PSII was performed, the diode leakage always increases with ideality factor. The samples with lower anneal temperatures are grouped toward higher leakage and higher ideality factor, suggesting that lower anneal temperatures result in an increased presence of generation-recombination (G-R) centers in the diode depletion region. Previously published ITS SIMS data show a reduction in N_{peak} with annealing temperature [14], [15]. This leads to a larger depletion region within the SDE, which exposes this depletion region to a higher Ni concentration, which increases G-R leakage [25], [31], [32]. Thus the increase in leakage with reduced annealing temperature is dominated by this effect as opposed to a reduction in minority carrier ϕ_B.

Additionally, with F-PSII and the same thermal budget, $X_{j,SDE}$ is smaller, which results in an increasing fraction of the depletion region existing in the lightly doped Si substrate and a smaller fraction existing in the SDE region (which correlates to the spatial Ni distribution). Thus, it can be concluded that the effect of F-PSII to reduce $X_{j,SDE}$ also reduces diode leakage, at least until $X_{j,SDE}$ becomes small enough for tunneling through the SDE region to dominate the diode leakage current.

Based on the presented data, which demonstrate a reduction in $X_{j,SDE}$ with F-PSII, it stands to reason that DSS junctions can be tuned by varying the F-PSII dose for a given silicide, ITS dose and energy, and post-ITS anneal time and temperature. Thus, a worthwhile future study would be to compare ITS for NiSi and PtSi and a range of F-PSII doses, to determine whether the lower limit for $X_{j,SDE}$ (and other diode characteristics, such as ϕ_B, leakage, etc.) for NiSi is indeed competitive with that for PtSi. Such a study would also provide insight into process optimization windows for DSS MOSFETs using different silicide materials.

6.6 DSS MOSFET Fabrication

The starting substrates were lightly doped p-type 6" silicon-on-insulator (SOI) wafers, with a 200 nm buried oxide (BOX) layer. The body thickness t_{body} was thinned to 30 nm by dry oxidation at 900 °C. After mesa active region patterning, a 6.5 nm (2.8 nm) NMOS (PMOS) dry gate oxide was grown, followed by 100 nm n^+ poly-$Si_{0.84}Ge_{0.16}$ gate LPCVD and then a 20 nm low-temperature-oxide (LTO) layer. The gate electrodes were defined using i-line lithography with photoresist trimming in O_2 plasma. After the gate stack was etched, the wafer was subjected to dry re-oxidation (to grow ~6.5 nm SiO_2) followed by 21 nm Si_3N_4 LPCVD. A masked F-PSII was then performed with F^+ at 20 keV, 0° tilt angle, and $1x10^{15}$ cm^{-2} dose (Fig. 6.10(a)). No post-implant anneal was performed, in order to maximize F segregation during the subsequent silicidation process [32]. After etching the Si_3N_4/SiO_2 gate-sidewall spacers, a 3 nm sputter pre-clean was performed, followed by 16 nm Ni sputter deposition. The wafer was then annealed at 300 °C/5 min in an oven with N_2 ambient. The unreacted Ni was then removed in a heated H_2SO_4/H_2O_2 solution and then NiSi was formed by rapid thermal annealing (RTA) at 500 °C/1 min (Fig. 6.10(b)). ITS was then performed with P^+ (20 keV) or BF_2 (25 keV) at 0° tilt angle, and $1x10^{15}$ cm^{-2} dose (Fig. 6.10(c)). The post-ITS anneal was performed by RTA in N_2 ambient at 600 °C for varying times to form n^+ SDE regions (Fig. 6.10(d)). The implant conditions were selected based on SRIM simulations [23], to provide for vertically uniform F-PSII, P-ITS, and B-ITS distributions within the thin SOI/silicide layer. The gate oxide was grown relatively thick, to amplify the impact of any change in SDE junction depth (due to F-PSII) as a change in short channel effects (SCE). Fig. 6.10(d) shows a schematic

cross-section of the fabricated device.

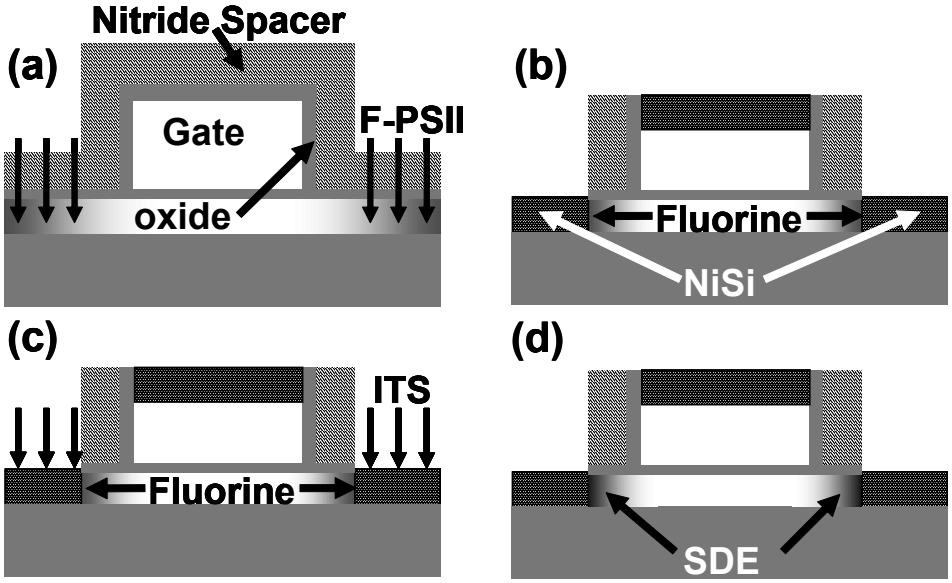

Fig. 6.10. Schematic cross-sections illustrating key steps in the DSS MOSFET fabrication process. (a) F-PSII is performed before (b) gate-sidewall spacers are etched and NiSi is formed, resulting in segregated F at the NiSi/Si interface. Then (c) P-ITS is performed and (d) the wafer is annealed at 600 °C for 1 min to form SDE regions.

6.6 DSS MOSFET Electrical Results

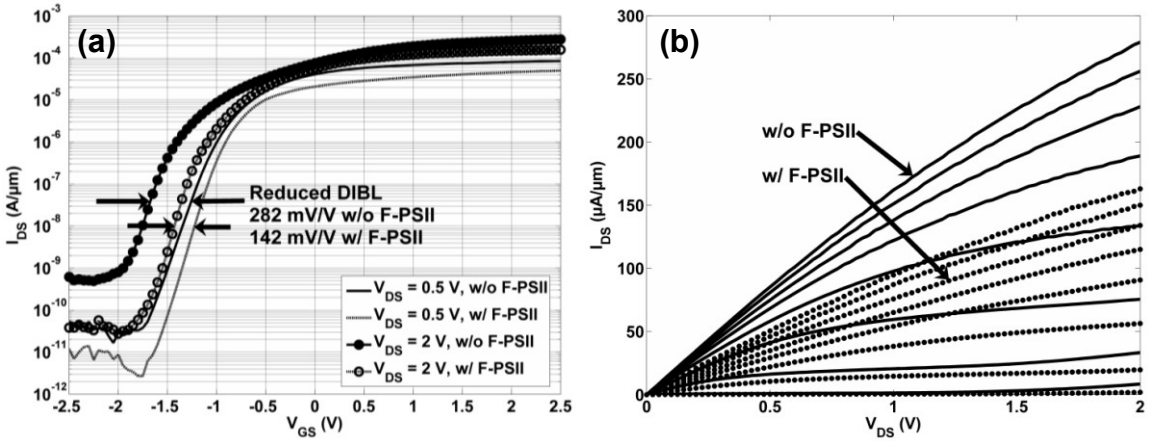

Fig. 6.11. Measured (a) I_{DS} vs. V_{GS} characteristics and (b) I_{DS} vs. V_{DS} characteristics for n-channel DSS NMOSFETs (L_G = 160 nm, W = 10 μm) fabricated with or without F-PSII. For (b), V_{GS} is varied from -1 V to 2.5 V in 0.5 V increments. The post-ITS anneal for these samples is 600 ºC for 1 min.

Fig. 6.11(a) shows I_{DS} vs. V_{GS} curves for DSS NMOSFETs with gate length L_G = 160 nm and channel width W = 10 μm. A reduction in drain-induced barrier lowering (DIBL) with F-PSII is clearly seen, from 282 mV/V to 142mV/V at 100 nA/μm, in Fig. 6.11(a), and in Fig. 6.12(a) over a range of L_G values. This indicates an increase in the effective channel length due to a reduction in $X_{j,SDE}$ with F-PSII, which also accounts for the lower I_{DS} at high V_{GS}. The I_{DS} vs. V_{DS} curves in

Fig. 6.11(b) show higher total resistance R_{total} in the linear region, which implies higher source/drain resistance R_{SD}, for the device with F-PSII. At V_{GS} = 2.5 V and V_{DS} = 50 mV, R_{total} increases from 5.85 kΩ-μm without F-PSII to 9.73 kΩ-μm with F-PSII. The NiSi sheet resistance R_s, although different between the two splits (8.80 ± 5.09 Ω/sq without F-PSII vs. 11.12 ± 4.34 Ω/sq with F-PSII, corresponding to 0.264 Ω-μm and 0.334 Ω-μm, respectively), represents a small fraction (~$4x10^{-5}$) of R_{total}. The linearity of the I_{DS} vs. V_{DS} curves at low V_{DS} suggests no significant increase in electron SBH, and therefore contact resistance, with F-PSII. Furthermore, F-PSII does not affect the silicide thickness [31], [32] (also supported by Fig. 6.2, where the segregated dopant peak position at ~25 nm does not change with F-PSII), meaning there should not be any difference in lateral silicidation between the devices with and without F-PSII. This all suggests that $X_{j,SDE}$ is the dominant factor in modulating R_{total} and, with F-PSII, is small enough to be contained largely or entirely under the gate-sidewall spacer, such that the SDE is gate-underlapped. The results herein do not imply that F-PSII would result in large R_{SD} in an optimally designed device. (The gate-sidewall spacer width can be reduced to lower R_{SD}.) The purpose of this study is solely to demonstrate that $X_{j,SDE}$, hence SCE, can be reduced by F-PSII, and this is clearly demonstrated in Figs. 6.11 and 6.12.

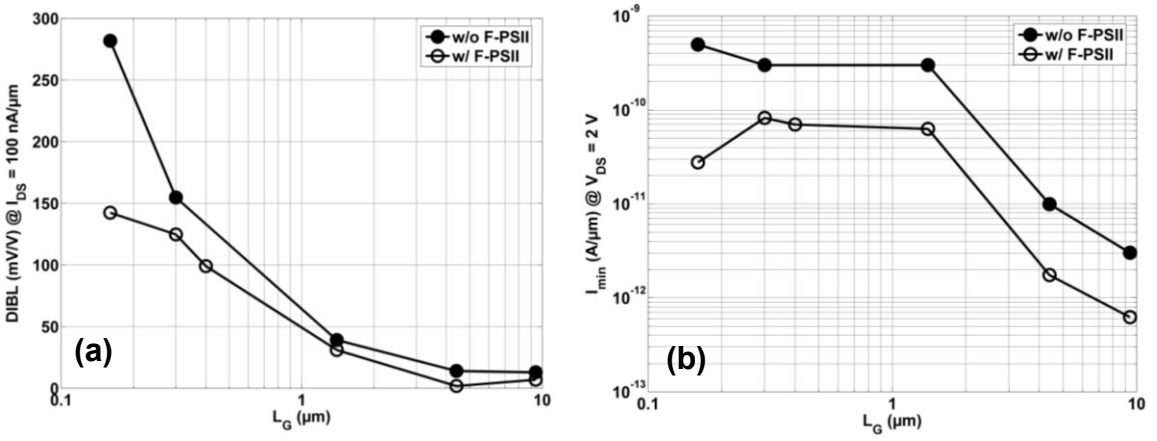

Fig. 6.12. Measured (a) DIBL vs. L_G and (b) I_{min} vs. L_G for n-channel DSS MOSFETs (W = 10 μm) fabricated with or without F-PSII. The post-ITS anneal for these samples is 600 ºC for 1 min.

As the anneal time increases, DIBL increases, even when F-PSII is utilized (Fig. 6.13). This degradation in DIBL is traceable to the reduction in effective channel length L_{eff} as $X_{j,SDE}$ increases with anneal time, as inferred from Fig. 6.5. Interestingly, for narrow devices (W = 0.7 μm) at the same L_G, DIBL is considerably lower, as shown in Fig. 6.13 for the devices with F-PSII. This is not due to a significance of the corner effect or reverse narrow width effect on improving gate control, since the active area aspect ratio remains very small at W = 0.7 μm (t_{body}/W ~ 0.043). (In other words, the "narrow" width devices by no means resemble a FinFET or even a Tri-Gate MOSFET.) Thus, the smaller degradation in device performance for narrow widths and long anneal times is due to larger L_{eff}, perhaps because the growth in $X_{j,SDE}$ with anneal time is smaller for narrow devices.

This effect of improved DIBL for narrower (but still planar) devices may be traced back to the NiSi grain size. Although the grain size for the samples measured here are unknown, grain sizes on the order of 200-300 nm have been reported in the literature for NiSi [31], [32], [47], [48]. So, to a first order, it is plausible to assume similar NiSi grain sizes for the presented work. Thus, for W = 0.7 μm, it is possible that there are only 3-4 NiSi grains interfacing the Si channel

region, as opposed to the case for W = 10 μm where there are 50+ NiSi grains. As mentioned previously in Section 6.3, in order for the dopants within the NiSi to diffuse into the adjacent Si, the NiSi/Si interface energy for a particular grain must be high enough (*i.e.*, low thermal stability, such as NiSi grains with (202) or (211) planes parallel to a (100) Si interface [26]) for that particular grain to begin rejecting Ni atoms. As the device width shrinks to a few NiSi grains, substantial dopant diffusion into the adjacent Si now becomes a function of the probability that any one of these grains has a high enough interface energy to reject Ni atoms (or more accurately, the probability that extent of Ni atom rejection is high, since the interface energy is always finite), as opposed to wide devices where the larger grain count interfacing the Si channel assures that at least one NiSi grain will have a high interface energy.

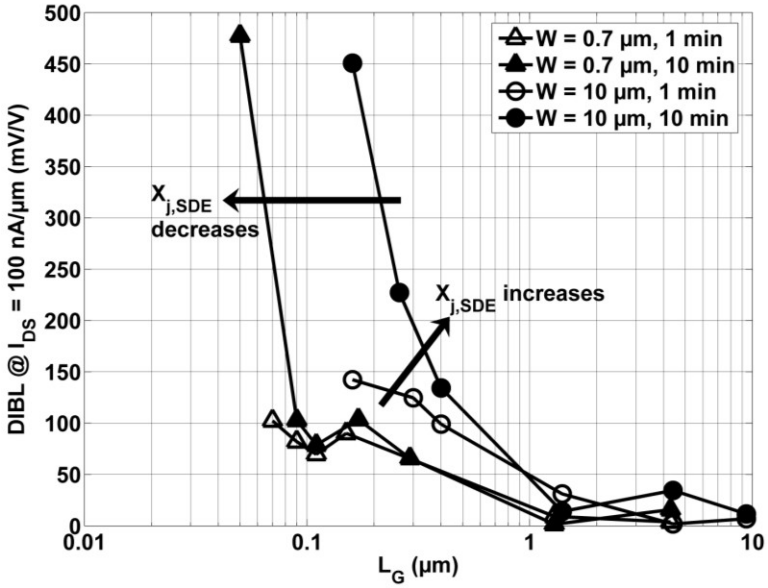

Fig. 6.13. Measured DIBL vs. L_G for n-channel DSS MOSFETs with F-PSII, for different device widths and post-ITS anneal times at 600 ºC.

Furthermore, it is possible that preferential NiSi grain orientation arises for narrow active regions, giving rise to lower average NiSi/Si interface energy (*e.g.*, NiSi grains with (013) or (020) planes parallel to a (100) Si interface [26]). This has not yet been proven or disproven, though, because x-ray diffraction (XRD), which reveals information about grain orientation [26], is a bulk measurement technique and so is not suitable for measuring narrow patterned lines. Yet another possibility is that L_c increases as the NiSi width decreases. Nonetheless, a detailed understanding of the effect of linewidth on NiSi thermal stability is a worthwhile future study and may lend more insight to the behavior shown in Fig. 6.13. This could be investigated by electron diffraction measurements scanned over the length of a narrow NiSi line (in order to sample multiple grains in-line), or by directly measuring $X_{j,SDE}$ for narrow vs. wide devices using high resolution scanning spreading resistance microscopy [50] (although, extracted dopant profiles would need to be calibrated against the effect of interstitial Ni on carrier mobility).

If the line of reasoning discussed here holds up against future experimentation, it may also explain why, in [3], the authors were able to demonstrate 25 nm NiSi DSS FinFETs using ITS and a substantial 600 ºC, 30min anneal, while still demonstrating reasonable control over short channel effects (SCE), since the fin width was small enough (40 nm) such that only one NiSi grain interfaces the Si channel region [49].

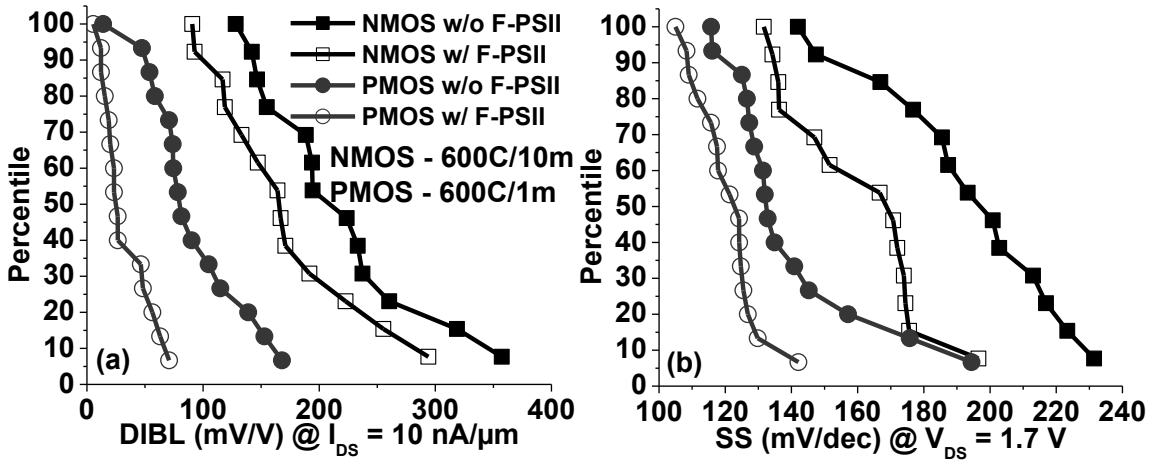

Fig. 6.14. (a) DIBL and (b) SS cumulative distributions. L/W = 50 nm/0.7 μm (NMOS) or 410 nm/10 μm (PMOS).

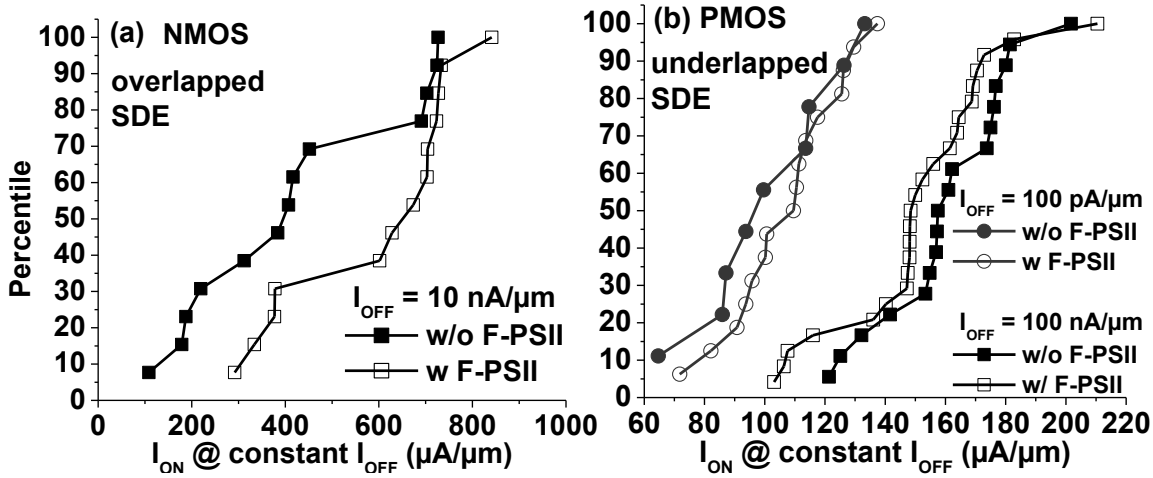

Fig. 6.15. Cumulative distributions of I_{ON} at constant I_{OFF} for (a) NMOS (L/W = 50 nm/0.7 μm, 600 ºC/10min anneal) and (b) PMOS (L/W = 410 nm/10 μm, 600 ºC/1min anneal). V_{DS} = 1.7 V and $(V_{GS} - V_{OFF})/T_{ox}$ ~ 6 MV/cm (4 V for NMOS, 1.7 V for PMOS).

To assure that the effect of F-PSII is real and not simply a coincidence for the data collected thus far (which came from two die on an entire wafer), statistical data was collected for the smallest yielding NMOS (600 ºC, 10 min anneal) and PMOS (600 ºC, 1 min anneal) devices. DIBL and subthreshold swing (SS) data are shown in Fig. 6.14, while I_{ON} at constant I_{OFF} is shown in Fig. 6.15. As Fig. 6.14 shows with statistical significance, short-channel behavior is considerably improved with F-PSII, since $X_{j,SDE}$ is reduced with F-PSII for both long (10 min) and short (1 min) post-ITS anneal times. With F-PSII, I_{ON} is higher for NMOS devices, due to reduced $X_{j,SDE}$ and improved short-channel control (Fig. 6.15(a)). For PMOS devices, (Fig. 6.15(b)) the effect of F-PSII on I_{ON} depends on the off-state leakage specification (I_{OFF}): it is higher for low I_{OFF} but lower for high I_{OFF}. This is due to the shorter (1 min) post-ITS anneal used for the PMOS devices, which resulted in a gate-underlapped structure with higher R_{SD}, and is similar to the NMOS 1 min anneal case from Fig. 6.11(a), where I_{ON} at I_{OFF} = 100 nA/μm is also lower for the F-PSII split (275.8 μA/μm without F-PSII vs. 161.1 μA/μm with F-PSII, at V_{DS} = 2 V and $(V_{GS} - V_{OFF})$ = 4 V). Again, the gate sidewall spacer can be reduced to offset this

effect, since the sidewall spacer thickness used here (~ 27 nm) was likely too large to be optimal for the resulting $X_{j,SDE}$ values after the short post-ITS anneal.

Based on the presented data, which demonstrate a reduction in $X_{j,SDE}$ with F-PSII, it stands to reason that DSS junctions can be tuned by varying the F-PSII dose and/or implant angle for a given silicide, ITS dose and energy, and post-ITS anneal time and temperature. (However, this must be balanced against F-PSII implant damage at the NiSi/Si interface which, if too high for heavy doses, will actually increase $X_{j,SDE}$.) Thus, a worthwhile future study would be to compare ITS for NiSi and PtSi and a range of F-PSII doses, to determine whether the lower limit for $X_{j,SDE}$ (and other diode characteristics, such as ϕ_B, leakage, etc.) for NiSi can be truly competitive with that for PtSi. The data presented here looks promising in this regard, and points toward single silicide DSS CMOS being the optimal integration scheme for metallic source/drain technology, whereby NMOS and PMOS devices can be individually tuned with separate F-PSII processes. Additional optimization paths include damage-less pre-silicidation treatment, such as N co-plasma [35] during Ni sputtering to improve the NiSi thermal stability, as well as pulsed laser annealing [51], to minimize the ramp thermal budget for the post-ITS anneal.

6.8 Summary

A fundamental understanding has been gained of the process by which DSS junctions are formed when ITS processing is utilized to form the junctions. At the most basic level, the formation of DSS junctions relies upon the silicide material having finite thermal stability, such that the silicide reaches a lower energy state during the post-ITS anneal by rejecting metal atoms. These metal atoms diffuse into the Si, giving rise to vacancies which greatly enhance dopant diffusion in Si at low temperatures, thus allowing the formation of the DSS junction. At some distance away from the silicide, the metal atoms form clusters and the vacancy concentration in the substrate increases considerably beyond this point. For sufficiently long post-ITS anneal times, this results in a kink in the DSS junction profile. By performing F-PSII, the thermal stability of the silicide can be improved, leading to less metal rejection from the silicide and therefore sharper DSS junctions. This also leads to two side effects, the first being a reduction in DSS diode leakage for the same post-ITS anneal temperature and the second being an improvement in the efficiency with which the majority carrier SBH can be reduced by the segregated dopants at the silicide/silicon interface. This leads to zero or near-zero SBH contacts over a wide anneal temperature range. (In this study, NiSi was used for ITS processing, but the physical mechanisms observed in this work should equally apply to other silicide materials.) As a result, F-PSII considerably improves the process design space for forming SDE regions in aggressively scaled DSS MOSFETs, as demonstrated by reduced DIBL and SS with F-PSII, for the MOSFETs fabricated in this study.

6.9 References

[1] G. Larrieu, E. Dubois, R. Valentin, N. Breil, F. Danneville, G. Dambrine, J. P. Raskin, J. C. Pesant," Low Temperature Implementation of Dopant-Segregated Band-edge Metallic S/D junctions in Thin-Body SOI p-MOSFETs," *IEDM Tech. Dig.*, pp. 147-150, 2007.

[2] 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 Tech. Dig.*, pp. 168-169, 2004.

[3] 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, pp. 430-432, Jun. 2004.

[4] V. Gudmundsson, P.-E. Hellstrom, J. Luo, J. Lu, S.-L. Zhang, M. Ostling, "Fully Depleted UTB and Trigate N-Channel MOSFETs Featuring Low-Temperature PtSi Schottky-Barrier Contacts With Dopant Segregation," *IEEE Elec. Dev. Lett.*, vol. 30, no. 5, pp. 541-543, May 2009.

[5] M. Awano, H. Onoda, K. Miyashita, K. Adachi, Y. Kawase, K. Miyano, H. Yoshimura, T. Nakayama, "Advanced DSS MOSFET Technology for Ultrahigh Performance Applications," *VLSI Tech. Dig.*, pp. 24-25, 2008.

[6] R. T. P. Lee, T.-Y. Liow, K.-M. Tan, A. E.-J. Lim, C.-S. Ho, K.-M. How, M. Y. Lai, T. Osipowicz, G.-Q. Lo, G. Samudra, D.-Z. Chi, Y.-C. Yeo, "Novel Epitaxial Nickel Aluminide-Silicide with Low Schottky-Barrier and Series Resistance for Enhanced Performance of Dopant-Segregated Source/Drain N-channel MuGFETs," *VLSI Tech. Dig.*, pp. 108-109, 2007.

[7] H.-W. Chen, C.-H. Ko, T.-J. Wang, C.-H. Ge, K. Wu, W.-C. Lee, "Enhanced Performance of Strained CMOSFETs Using Metallized Source/Drain Extension (M-SDE)," *VLSI Tech. Dig.*, pp. 118-119, 2007.

[8] J. Knoch, M. Zhang, Q. T. Zhao, St. Lenk, S. Mantl, J. Appenzeller, "Effective Schottky barrier lowering in silicon-on-insulator Schottky-barrier metal-oxide-semiconductor field-effect transistors using dopant segregation," *Appl. Phys. Lett.*, vol. 87, p. 263505, 2005.

[9] R. A. Vega, "Schottky field effect transistors and Schottky CMOS circuitry," M.S. Thesis, Dept. Microelectron. Eng., Rochester Inst. Technol., Rochester, NY, 2006.

[10] L. Hutin, M. Vinet, T. Poiroux, C. Le Royer, B. Previtali, C. Vizioz, D. Lafond, Y. Morand, M. Rivoire, F. Nemouchi, V. Carron, T. Billon, S. Deleonibus, O. Faynot, "Dual Metallic Source and Drain Integration on Planar Single and Double Gate SOI CMOS down to 20nm: Performance and Scalability Assessment," *IEDM Tech. Dig.*, pp. 56-59, 2009.

[11] G. Larrieu, D. A. Yarekha, E. Dubois, N. Breil, O. Faynot, "Arsenic-Segregated Rare-Earth Silicide Junctions: Reduction of Schottky Barrier and Integration in Metallic n-MOSFETs on SOI," *IEEE Elec. Dev. Lett.*, vol. 30, no. 12, pp. 1266-1268, Dec. 2009.

[12] A. Kaneko, A. Yagashita, K. Yahashi, T. Kubota, M. Omura, K. Matsuo, I. Mizushima, K. Okano, H. Kawasaki, T. Izumida, T. Kanemura, N. Aoki, A. Kinoshita, J. Koga, S. Inaba, K. Ishimaru, Y. Toyoshima, H. Ishiuchi, K. Suguro, K. Eguchi, Y. Tsunashima, "High-Performance FinFET with Dopant-Segregated Schottky Source/Drain," *IEDM Tech. Dig.*, pp. 893-896, 2006.

[13] V. Carron, F. Nemouchi, Y. Morand, T. Poiroux, M. Vinet, S. Bernasconi, O. Loveau, D. Lafond, V. Delaye, F. Allain, S. Minoret, L. Vandroux, T. Billon, "Platinum Silicide Metallic Source & Drain Process Optimization for FDSOI pMOSFETs, " *IEEE International SOI Conference*, pp. 111-112, 2009.

[14] T. Yamauchi, Y. Nishi, Y. Tsuchiya, A. Kinoshita, J. Koga, K. Kato, "Novel doping technology for a 1nm NiSi/Si junction with dipoles comforting Schottky (DCS) barrier," *IEDM Tech. Dig.*, pp. 963-966, 2007.

[15] Z. Qiu, Z. Zhang, M. Ostling, S.-L. Zhang, "A Comparative Study of Two Different Schemes to Dopant Segregation at NiSi/Si and PtSi/Si interfaces for Schottky Barrier Height Lowering," *IEEE Trans. Elec. Dev.*, vol. 55, no. 1, pp. 396-403, Jan. 2008.

[16] C.-H. Shih, S.-P. Yeh, "Device considerations and design optimizations for dopant segregated Schottky barrier MOSFETs," *Semicond. Sci. Tec.*, vol. 23, p. 125033, 2008.

[17] R. A. Vega, T.-J. King Liu, "A comparative study of dopant-segregated Schottky and raised source/drain double-gate MOSFETs," *IEEE Trans. Elec. Dev.*, vol. 55, no. 10, pp. 2665-2677, Oct. 2008.

[18] R. A. Vega, K. Liu, T.-J. King Liu, "Dopant-Segregated Schottky Source/Drain Double-Gate MOSFET Design in the Direct Source-to-Drain Tunneling Regime," *IEEE Trans. Elec. Dev.*, vol. 56, no. 9, pp. 2016-2026, Sept. 2009.

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

[20] C. L. Chu, K. C. Saraswat, S. S. Wong, "Characterization of Lateral Dopant Diffusion in Silicides," *IEDM Tech. Dig.*, pp. 245-248, 1990.

[21] 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, pp. 4619-4626, Nov. 1992.

[22] M. Wittmer, K. N. Tu, "Low-temperature diffusion of dopant atoms in silicon during interfacial silicide formation," *Phys. Rev. B*, vol. 29, no. 4, pp. 2010-2020, Feb. 1984.

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

[24] J. A. Kittl, K. Opsomer, C. Torregiani, C. Demeurisse, S. Mertens, D. P. Brunco, M. J. H. Van Dal, A. Lauwers, "Silicides and germanides for nano-CMOS applications," *Mat. Sci. Eng. B*, vol. 154-155, pp. 144-154, 2008.

[25] M. Tsuchiaki, K. Ohuchi, C. Hongo, "Junction Leakage Generation by NiSi Thermal Instability Characterized Using Damage-Free n+/p Silicon Diodes," *Jpn. J. Appl. Phys*, vol. 43, no. 8A, pp. 5166-5173, 2004.

[26] D. Deduytsche, C. Detavernier, R. L. Van Mierhaeghe, C. Lavoie, "High-temperature degradation of NiSi films: Agglomeration versus $NiSi_2$ nucleation," *J. Appl. Phys.*, vol. 98, p. 033526, 2005.

[27] K. Ohuchi, C. Lavoie, C. Murray, C. D'Emic, I. Lauer, J. O. Chu, B. Yang, P. Besser, L. Gignac, J. Bruley, G. U. Singco, F. Pagetter, A. W. Topol, M. J. Rooks, J. J. Bucchignano, V. Narayanan, M. Khare, M. Takayanagi, K. Ishimaru, D.-G. Park, G. Shahidi, P. Solomin, "Extendibility of NiPt Silicide Contacts for CMOS Technology Demonstrated to the 22-nm Node," *IEDM Tech. Dig.*, pp. 1029-1031, 2007.

[28] T. Marukame, T. Yamauchi, Y. Nishi, T. Sasaki, A. Kinoshita, J. Koga, K. Kato, "Impact of platinum incorporation on thermal stability and interface resistance in NiSi/Si junctions based on first-principles calculation," *IEDM Tech. Dig.*, pp. 547-550, 2008.

[29] T. Sonehara, A. Hokazono, H. Akutsu, T. Sasaki, H. Uchida, M. Tomita, H. Tsujii, S. Kawanaka, S. Inaba, Y. Toyoshima, "Contact resistance reduction of Pt-incorporated NiSi for continuous CMOS scaling ~ Atomic level analysis of Pt/B/As distribution within silicide films ~," *IEDM Tech. Dig.*, pp. 921-924, 2008.

[30] C.-C. Wang, M.-C. Chen, "Formation and Characterization of NiSi-Silicided n+p Shallow Junctions," *Jpn. J. Appl. Phys.*, vol. 45, no. 3A, pp. 1582-1587, 2006.

[31] M. Tsuchiaki, K. Ohuchi, A. Nishiyama, "Suppression of Thermally Induced Leakage of NiSi-Silicided Shallow Junctions by Pre-Silicide Fluorine Implantation," *Jpn. J. Appl. Phys.*, vol. 44, no. 4A, pp. 1673-1681, 2005.

[32] M. Tsuchiaki, A. Nishiyama, "Substrate Orientation Dependent Suppression of NiSi Induced Junction Leakage by Fluorine and Nitrogen Incorporation," *Jpn. J. Appl. Phys.*, vol. 47, no. 4, pp. 2388-2397, 2008.

[33] W.-Y. Loh, P. Y. Hung, B. E. Coss, P. Kalra, I. Ok, G. Smith, C.-Y. Kang, S.-H. Lee, J. Oh, B. Sassman, P. Majhi, P. Kirsch, H.-H. Tseng, R. Jammy, "Selective Phase Modulation of NiSi Using N-ion implantation for High Performance Dopant-Segregated Source/Drain n-Channel MOSFETs," *VLSI Tech. Dig.*, pp. 100-101, 2009.

[34] D. Diebel, S. Chakravarthi, S. T. Dunham, C. F. Machala, S. Ekbote, A. Jain, "Investigation and modeling of fluorine co-implantation effects on dopant redistribution," *Mat. Res. Soc. Symp. Proc.*, vol. 765, pp. D.6.15.1-D.6.15.6., 2003.

[35] B. Imbert, M. Gregoire, S. Zoll, R. Beneyton, S. Del-Medico, C. Trouiller, O. Thomas, "Nitrogen impurity effects in nickel silicide formation at low temperatures – New "nitrogen co-plasma" approach," *Microelectronic Engineering*, vol. 85, pp. 2005-2008, 2008.

[36] N. Stavitski, M. J. H. van Dal, A. Lauwers, C. Vrancken, A. Y. Kovalign, R. A. M. Wolters, "Systematic TLM measurements of NiSi and PtSi specific contact resistance to n- and p-type Si in a broad doping range," *IEEE Elec. Dev. Lett.*, vol. 29, no. 4, pp. 378-381, Apr. 2008.

[37] N. Stavitski, M. J. H. van Dal, A. Lauwers, C. Vrancken, A. Y. Kovalign, R. A. M. Wolters, "Evaluation of transmission line model structures for silicide-to-silicon specific contact resistance extraction," *IEEE Trans. Elec. Dev.*, vol. 55, no. 5, pp. 1170-1176, May 2008.

[38] J. A. Kittl, K. Opsomer, C. Torregiani, C. Demeurisse, S. Mertens, D. P. Brunco, M. J. H. van Dal, A. Lauwers, "Silicides and germanides for nano-CMOS applications," *Mat. Sci. Eng. B*, vol.154-155, pp. 144-154, 2008.

[39] O. Chamirian, J. A. Kittl, A. Lauwers, O. Richard, M. van Dal, K. Maex, "Thickness scaling issues of Ni silicide," *Microelectronic Engineering*, vol. 70, pp. 201-208, 2003.

[40] H. F. Hsu, C. L. Tsai, H. Y. Chan, T. H. Chen, "Effect of P addition on the thermal stability and electrical characteristics of NiSi films," *Thin Solid Films*, vol. 518, pp. 1538-1542, 2009.

[41] D. K. Shroeder, *Semiconductor Material and Device Characterization*, 3rd ed. Hoboken, NJ: Wiley, 2006, pp. 161-162.

[42] S. Tanaka, K. Matsushita, H. Kitagawa, "Majority-Carrier Capture Cross Section of Amphoteric Nickel Center in Silicon Studied by Isothermal Capacitance Transient Spectroscopy," *Jpn. J. Appl. Phys.*, vol. 35, pp. 4624-4625, 1996.

[43] P. Kalra, "Advanced Source/Drain Technologies for Nanoscale CMOS," Ph.D. Thesis, Dept. Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA, 2008.

[44] E. J. Tan, K. L. Pey, D. Z. Chi, P. S. Lee, L. J. Tang, "Improved Electrical Performance of Erbium Silicide Schottky Diodes Formed by Pre-RTA Amorphization of Si," *IEEE Elec. Dev. Lett.*, vol. 27, no. 2, pp. 93-95, Feb. 2006.

[45] S. C. Jain, D. J. Roulston, "A Simple Expression for Band Gap Narrowing (BGN) in Heavily Doped Si, Ge, GaAs and Ge_xSi_{1-x} Strained Layers," *Solid-State Electronics*, vol. 34, no. 5, pp. 453-465, 1991.

[46] R. A. Vega, V. C. Lee, T.-J. King Liu, "The Effect of Random Dopant Fluctuation on Specific Contact Resistivity," *IEEE Trans. Elec. Dev.*, vol. 57, no. 1, pp. 273-281, Jan. 2010.

[47] S. K. Donthu, D. Z. Chi, S. Tripathy, A. S. W. Wong, S. J. Chua, "Micro-Raman spectroscopic investigation of NiSi films formed on BF_2^+-, B^+- and non-implanted (100)Si substrates," *Appl. Phys. A*, vol. 79, pp. 637-642, 2004.

[48] A. S. W. Wong, D. Z. Chi, M. Loomans, D. Ma, M. Y. Lai, W. C. Tjiu, s. J. Chua, C. W. Lim, J. E. Greene, "F-enhanced morphological and thermal stability of NiSi films on BF_2^+-implanted Si(001)," *Appl. Phys. Lett.*, vol. 81, no. 27, pp. 5138-5140, Dec. 2002.

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

[50] L. Zhang, M. Saitoh, A. Kinoshita, N. Yasutake, A. Hokazono, N. Aoki, N. Kusunoki, I. Mizushima, M. Koike, S. Takeno, J. Koga, "Insight into the S/D Engineering by High-resolution Imaging and Precise Probing of 2D-Carrier Profiles with Scanning Spreading Resistance Microscopy," *IEDM Tech. Dig.*, pp. 46-49, 2009.

[51] C. Ortolland, E. Rosseel, N.Horiguchi, C. Kerner, S. Mertens, J. Kittl, E. Verleysen, H. Bender, W. Vandervost, A. Lauwers, P.P. Absil, S. Biesemans, S. Muthukrishnan, S. Srinivasan, A.J. Mayur, R. Schreutelkamp, T. Hoffman, "Silicide Yield Improvement with NiPtSi Formation by Laser Anneal for Advanced Low Power Platform CMOS Technology," *IEDM Tech. Dig.*, pp. 23-26, 2009.

Chapter 7

Silicon Germanium Process Technology

7.1 Introduction

$Si_{1-x}Ge_x$ has found a number of applications in integrated circuits, due to its attractive electrical and physical properties. For example, polycrystalline $Si_{1-x}Ge_x$ (poly-$Si_{1-x}Ge_x$) is a favorable MOSFET gate material as compared against poly-Si, due to a higher degree of dopant activation resulting in reduced gate depletion effects [1], [2]. In addition, it mitigates the issues of gate Fermi-level pinning and gate leakage for integration of high-permittivity gate dielectrics [3]. Poly-$Si_{1-x}Ge_x$ is also a promising structural material for post-CMOS integration of micro-electro-mechanical devices (MEMS), because it can be formed at relatively low temperatures [4]. In consideration of the thermal budget limits for advanced CMOS devices/circuits, the $Si_{1-x}Ge_x$ deposition temperature should not exceed 430^{o}C to avoid degradation of interconnect (via) resistance [5].

Poly-$Si_{1-x}Ge_x$ films for gate-electrode and MEMS applications are typically formed by low pressure chemical vapor deposition (LPCVD). Thin epitaxial layers of $Si_{1-x}Ge_x$ can also be formed using low-temperature LPCVD, *e.g.* to provide heterostructure channels for enhanced MOSFET performance [6]. For these applications, it is desirable to be able to dope the film in a well-controlled manner (with respect to dopant concentration and Ge content in the deposited film, and film deposition rate). Previous works have investigated *in-situ* doping of $Si_{1-x}Ge_x$ films deposited at moderate-to-high temperatures (≥450°C) [7] – [10], with limited data for films deposited at low temperatures [9], [10]. This chapter investigates low-temperature (425^{o}C) deposition of *in-situ* doped polycrystalline $Si_{1-x}Ge_x$ films by LPCVD, to characterize the film deposition rate and dopant incorporation as a function of Si gas source (SiH_4 vs. Si_2H_6), Ge content, and dopant gas flow. Various deposition mechanisms are elucidated to help guide LPCVD process optimization, especially for n-type films since phosphorus poisoning is significant at low deposition temperatures [11]. Additionally, solid phase epitaxial regrowth (SPER) and Ge melt processing are investigated as methods to form crystalline $Si_{1-x}Ge_x$ out of LPCVD Ge and $Si_{1-x}Ge_x$ layers.

7.2 LPCVD of In-Situ Doped N- and P-Type $Si_{1-x}Ge_x$ at 425 ºC

Table 7.1. LPCVD Si deposition recipes

Recipe	Step #	Time (min)	Gas Flow Rates (sccm)						Partial Pressure (mTorr)	
			GeH_4	SiH_4	Si_2H_6	PH_3	BCl_3	Total	Dopant Gas	GeH_4
NS1	1	60	-	-	66	9	-	75	24	-
	2	60	-	-	68	7	-	75	18.67	-
	3	60	-	-	70	5	-	75	13.33	-
	4	60	-	-	72	3	-	75	8	-
	5	60	-	-	74	1	-	75	2.67	-
PS1	1	40	-	-	90	-	50	140	1.43	-
	2	40	-	-	100	-	40	140	1.14	-
	3	40	-	-	110	-	30	140	0.86	-
	4	40	-	-	120	-	20	140	0.57	-
	5	40	-	-	130	-	10	140	0.29	-
PS2	1	60	-	90	-	-	50	140	1.43	-
	2	60	-	100	-	-	40	140	1.14	-
	3	60	-	110	-	-	30	140	0.86	-
	4	60	-	120	-	-	20	140	0.57	-
	5	60	-	130	-	-	10	140	0.29	-
PS3	1	40	-	155	-	-	45	200	0.9	-
	2	40	-	165	-	-	35	200	0.7	-
	3	40	-	175	-	-	25	200	0.5	-
	4	40	-	185	-	-	15	200	0.3	-
	5	40	-	195	-	-	5	200	0.1	-

All depositions were performed on bare 150-mm-diameter Si (100) wafer substrates in a Tystar horizontal hot-wall LPCVD reactor with caged boats. Wafers were first cleaned in a H_2SO_4/H_2O_2 bath at 120 °C, followed by de-ionized water rinsing, before being loaded into the reactor tube. The boats were placed at the rear of the tube; the source gases were injected through a gas ring at the front of the tube. The temperature and pressure for all depositions was 425°C and 400 mTorr. GeH_4 was used as the Ge source gas, and either SiH_4 or Si_2H_6 was used as the Si source gas. BCl_3 (1% BCl_3 in 99% He) and PH_3 (50% PH_3 in 50% SiH_4) were used as the dopant gases for p-type and n-type films, respectively. Multi-layer deposition recipes were used to save time and analysis cost. For each layer, the total gas flow was kept the same; only the partial pressures of the source gases were varied from layer to layer. The recipes used for Si and $Si_{1-x}Ge_x$ deposition are shown in Tables 7.1 and 7.2, respectively. It is noted that the calculated partial pressures are based on the recipe parameters, which may differ from the actual partial pressures at and near the wafer boat due to precursor gas depletion from the gas ring to the wafers. In all cases, wafers from the same wafer boat slot (*i.e.*, same position within the tube) were used for analysis to minimize any confounding of precursor gas depletion effects with the other mechanisms characterized in this study.

Table 7.2. LPCVD $Si_{1-x}Ge_x$ deposition recipes

Recipe	Step #	Time (min)	Gas Flow Rates (sccm)						Partial Pressure (mTorr)	
			GeH_4	SiH_4	Si_2H_6	PH_3	BCl_3	Total	Dopant Gas	GeH_4
NSG1	1	20	10	87	-	3	-	100	6	40
	2	20	10	84	-	6	-	100	12	40
	3	20	10	80	-	10	-	100	20	40
NSG2	1	20	10	187	-	3	-	200	3	20
	2	20	10	184	-	6	-	200	6	20
	3	20	10	181	-	9	-	200	9	20
	4	20	15	182	-	3	-	200	3	30
	5	20	15	179	-	6	-	200	6	30
	6	20	15	176	-	9	-	200	9	30
	7	20	20	177	-	3	-	200	3	40
	8	20	20	174	-	6	-	200	6	40
	9	20	20	171	-	9	-	200	9	40
NSG3	1	30	10	-	187	3	-	200	3	20
	2	30	10	-	184	6	-	200	6	20
	3	30	10	-	181	9	-	200	9	20
	4	30	15	-	182	3	-	200	3	30
	5	30	15	-	179	6	-	200	6	30
	6	30	15	-	176	9	-	200	9	30
	7	30	20	-	177	3	-	200	3	40
	8	30	20	-	174	6	-	200	6	40
	9	30	20	-	171	9	-	200	9	40
NSG4	1	30	18	-	54	3	-	75	8	96
	2	30	18	-	51	6	-	75	16	96
	3	30	18	-	48	9	-	75	24	96
	4	30	27	-	45	3	-	75	8	144
	5	30	27	-	42	6	-	75	16	144
	6	30	27	-	39	9	-	75	24	144
	7	30	36	-	36	3	-	75	8	192
	8	30	36	-	33	6	-	75	16	192
	9	30	36	-	30	9	-	75	24	192
PSG1	1	20	10	185	-	-	5	200	0.1	20
	2	20	10	180	-	-	10	200	0.2	20
	3	20	10	175	-	-	15	200	0.3	20
	4	20	15	180	-	-	5	200	0.1	30
	5	20	15	175	-	-	10	200	0.2	30
	6	20	15	170	-	-	15	200	0.3	30
	7	20	20	175	-	-	5	200	0.1	40
	8	20	20	170	-	-	10	200	0.2	40
	9	20	20	165	-	-	15	200	0.3	40

Fig. 7.1 shows P concentration in the deposited film *vs.* PH_3 partial pressure for n-type Si and $Si_{1-x}Ge_x$ deposited using Si_2H_6 as the Si source gas. For small percentages of Ge (< 10 atomic percent) and low PH_3 partial pressures, P incorporation differs little between $Si_{1-x}Ge_x$ and Si. As PH_3 partial pressure increases, a slight dependence of P concentration on Ge content emerges. As the Ge content increases further, P incorporation increases dramatically. The dependence of P incorporation on PH_3 partial pressure does not change with Ge content for high PH_3 partial pressures, which suggests that Ge does not change the P adsorption rate, but instead increases P concentration through a combination of promoting P desorption and suppressing P-P dimer

formation, to increase the availability of adsorption sites and to increase the amount of monatomic P that is more readily incorporated into the film.

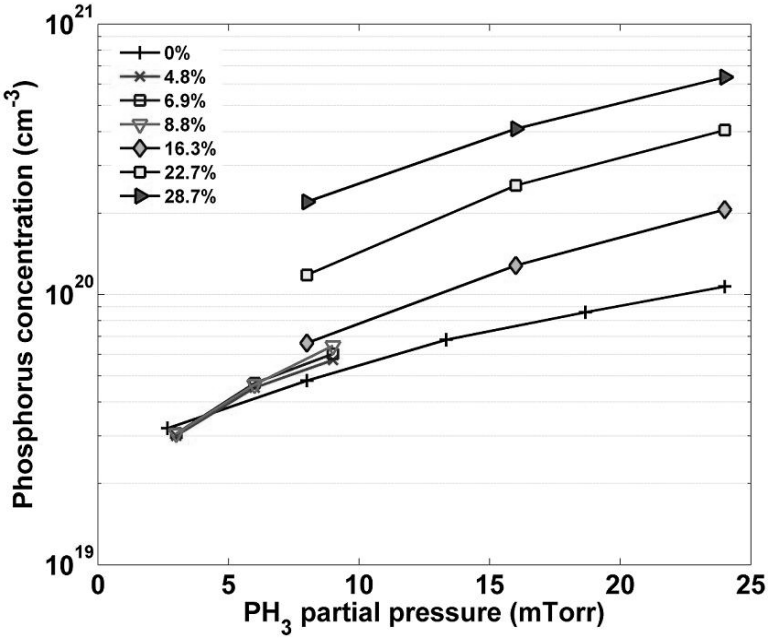

Fig. 7.1. P concentration vs. PH_3 partial pressure, for n-type films of various Ge content. Si_2H_6 was used as the Si source gas. Recipe NS1 (Table I) corresponds to the data with 0% Ge. Recipe NSG3 (Table II) corresponds to the data with 4.8-8.8% Ge, and recipe NSG4 corresponds to the data with 16.3-28.7% Ge.

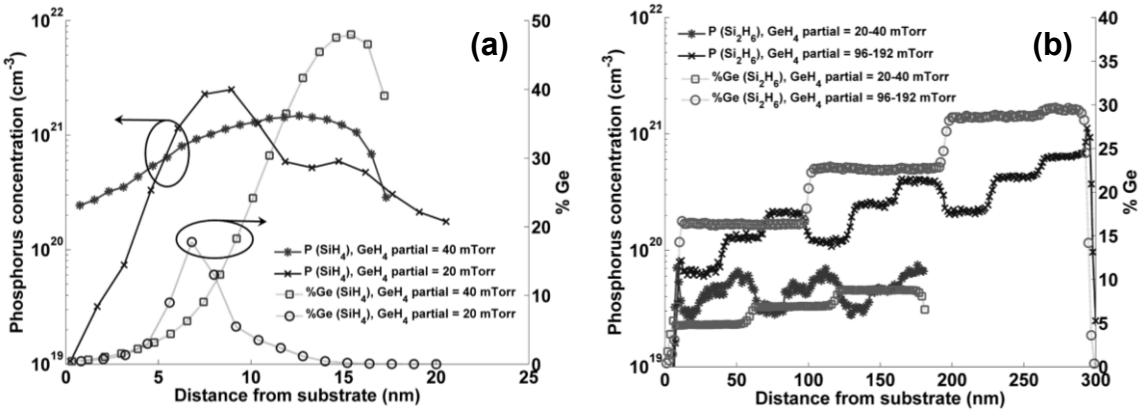

Fig. 7.2. P concentration and Ge content vs. distance from the substrate, for n-type $Si_{1-x}Ge_x$ multi-layer film stacks deposited using (a) SiH_4 and (b) Si_2H_6 as the Si source gas, for various PH_3 and GeH_4 partial pressures. In (a), the 40 mTorr case corresponds to recipe NSG1 (Table II), while the 20 mTorr case corresponds to recipe NSG2. In (b), the 20-40 mTorr case corresponds to recipe NSG3, while the 96-192 mTorr case corresponds to recipe NSG4.

These two effects compete with each other, as discussed in [11]. For high P surface coverage, P-P dimerization limits the number of available adsorption sites, which retards film deposition. A reduction in P-P dimers results in more available bonding sites around a P adatom, thus promoting film deposition. Ge-induced P desorption reduces the P concentration at the surface, but if the rate of P-P dimer suppression exceeds the rate of P desorption, there will be a net gain in the amount of P incorporated into the deposited film. This seems to be the case in Fig. 7.1 for moderate concentrations of Ge. For very low PH_3 partial pressures, however, P incorporation is

actually lower for $Si_{1-x}Ge_x$ than for Si. This suggests that the effect of Ge-induced P desorption dominates over P-P dimer suppression as the PH_3 partial pressure is reduced. This is consistent with the notion that fewer P-P dimers are formed when the surface coverage of P is reduced, since fewer P adatoms are likely to be adjacent to form a dimer.

P concentration and Ge concentration depth profiles for film stacks deposited using SiH_4 *vs.* Si_2H_6 as the Si source gas are shown in Fig. 7.2(a) *vs.* Fig. 7.2(b), respectively. The PH_3 partial pressure was varied from layer to layer for each of these film stacks. The GeH_4 partial pressure was also varied, for the film stack in Fig. 7.2(b). The P poisoning effect (suppression of film deposition) is manifest in the much thinner stack and poor control of P and Ge content in Fig. 7.2(a). This effect is essentially eliminated if Si_2H_6 is used as the Si source gas, as can be seen from Fig. 7.2(b). In contrast to previously reported findings for films deposited at higher temperatures [7], GeH_4 does not dramatically reduce P poisoning at low temperature with SiH_4 as the Si source gas. This may be due to the difference in the reactive sticking coefficient (RSC), which is ~10× higher for deposition using Si_2H_6 as compared with SiH_4 at low temperatures [12]. (The RSC is the probability that a species will adsorb to the substrate surface.) Also, the RSC of PH_3 was found to be greater than 40x that of SiH_4 [11], which results in highly preferential adsorption of PH_3 over SiH_4. Due to the lower RSC associated with SiH_4, more adsorption sites are available for Ge and P, which increases P surface coverage and therefore P-P dimerization. This results in a runaway process of progressively reduced SiH_4 adsorption due to P-P dimerization, and progressively increased P-P dimerization due to reduced SiH_4 adsorption, until the deposition ceases altogether. The much higher RSC associated with Si_2H_6 reduces P coverage, thereby suppressing the P-P dimerization that impedes deposition. It can be seen from Fig. 7.2(b) that, with the appropriate Si source gas, GeH_4 does help to mitigate P poisoning, since higher GeH_4 partial pressure results in a smoother P profile, *i.e.* improved control of P incorporation and film deposition rate.

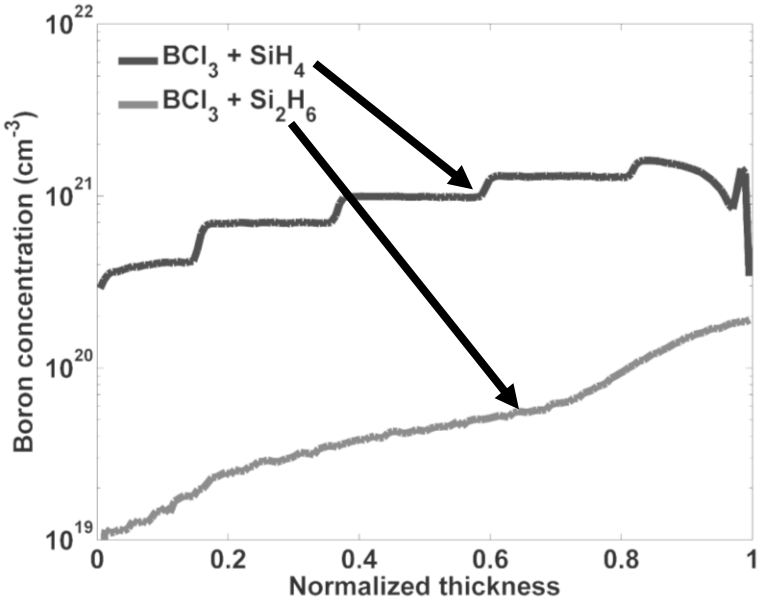

Fig. 7.3. B concentration vs. normalized distance from the substrate, for p-type Si film stacks deposited using SiH_4 and Si_2H_6 as the Si source gases. The Si_2H_6 case corresponds to recipe PS1 (Table I), while the SiH_4 case corresponds to recipe PS2.

For p-type films, well-controlled deposition is achieved with SiH_4 rather than Si_2H_6, as can be seen from Fig. 7.3, which compares the B concentration *vs.* normalized distance from the substrate for two Si film stacks. The BCl_3 partial pressure was varied from layer to layer for each of these film stacks; the deposition recipes used were the same, except for differences in the deposition times and the choice of Si source gas. The B concentration stabilizes more quickly upon a change in BCl_3 partial pressure, if SiH_4 is used as the Si source gas. B incorporation is not as well controlled, and dramatically reduced if Si_2H_6 is used instead of SiH_4. It should be noted that the reduction in B concentration is not attributable to an increased deposition rate for Si_2H_6-source deposition, because the deposition rate of p-type Si using Si_2H_6 (~1 nm/min) is actually lower than that using SiH_4 (~1.2-1.5 nm/min). Again, the difference in RSC between Si_2H_6 and SiH_4 can provide an explanation for this. Considering that there is a two-fold increase in the amount of Si provided by Si_2H_6 as compared with SiH_4, as well as the ~10x difference in RSC between both gases, for the same gas flow rates, the B:Si ratio for SiH_4-source deposition should be ~20× higher than for Si_2H_6-source deposition. The ratio of B concentrations for the film stacks in Fig. 7.3 is ~10× at higher B concentrations, and increases to ~ 27× at lower B concentrations, which supports this theory. The change in this ratio with B concentration suggests that the effect of B adatoms on SiH_4 adsorption is also significant. The higher deposition rate for SiH_4-source deposition is attributable to the higher B incorporation (due to the smaller RSC of SiH_4), which enhances SiH_4 adsorption due to an increase in potential at the deposition surface [13], resulting in a 20-50% increase in deposition rate. As for B concentration control and stabilization, Fig. 7.3 suggests that BCl_3 adsorbs less easily when adjacent to Si_2H_6, in comparison to SiH_4. This results in a self-feeding effect similar to P poisoning, but opposite in nature and less severe, which reduces the B concentration as the deposition proceeds.

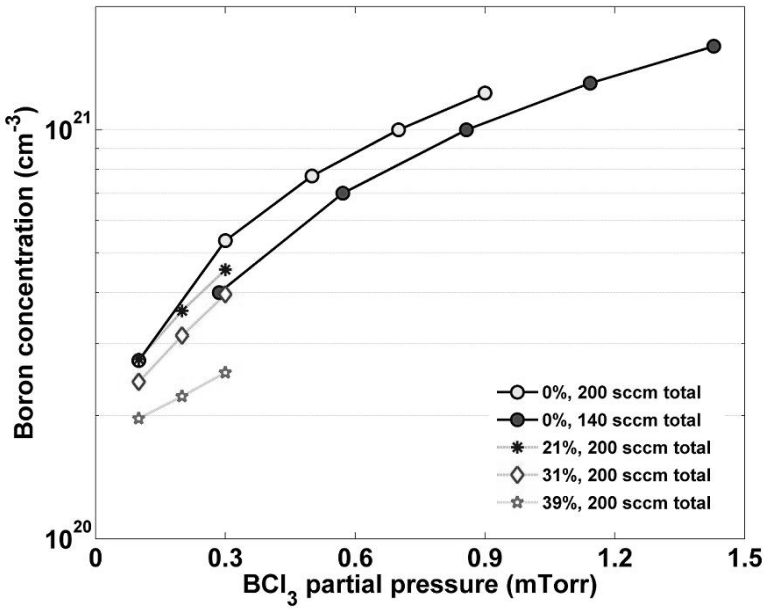

Fig. 7.4. B concentration vs. BCl_3 partial pressure for p-type films of various Ge content. SiH_4 was used as the Si source gas. For the 0% Ge, 140 sccm total and 200 sccm total cases, recipes PS2 and PS3 (Table I) were used, respectively. For the data with 21-39% Ge, recipe PSG1 (Table II) was used.

Fig. 7.4 shows how B concentration varies with BCl_3 partial pressure, for films deposited using SiH_4 as the Si source gas. For a given BCl_3 partial pressure, increased Ge content results in reduced B concentration. Also, the dependence of B concentration on BCl_3 partial pressure is weaker for films with higher Ge content. These results clearly indicate that GeH_4 more successfully competes for adsorption sites than BCl_3 and may in fact promote B desorption from the surface. (The effect of GeH_4 on B incorporation cannot be attributed to enhanced deposition rate, since the deposition rate was not found to vary significantly with Ge content.)

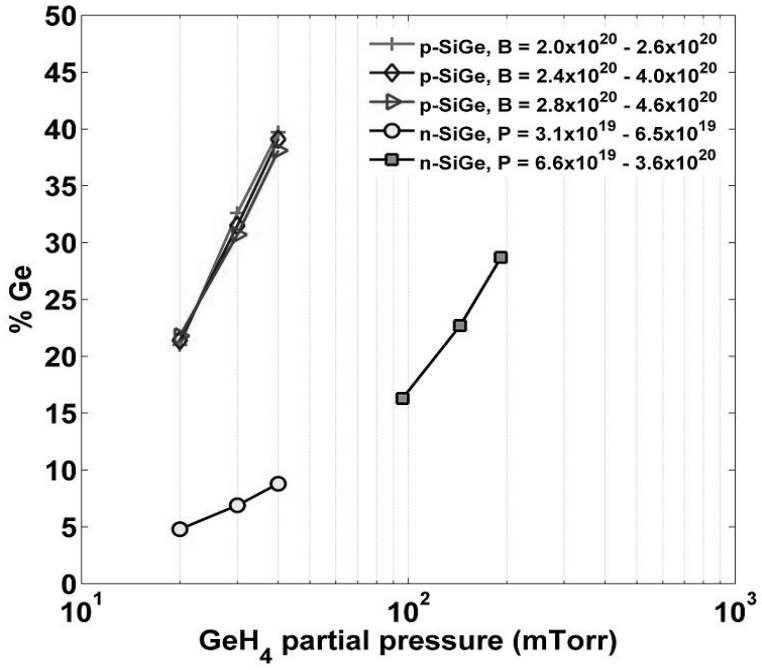

Fig. 7.5. Ge content in the deposited film *vs*. GeH_4 partial pressure during deposition, for n-type and p-type $Si_{1-x}Ge_x$. SiH_4 was used for p-type $Si_{1-x}Ge_x$ and Si_2H_6 was used for n-type $Si_{1-x}Ge_x$. For the p-SiGe data, recipe PSG1 (Table II) was used. The n-SiGe data at the lower GeH_4 partial pressure corresponds to recipe NSG3 (Table II), while the n-SiGe data at the higher GeH_4 partial pressure corresponds to recipe NSG4.

From the two curves for pure Si deposition in Fig. 7.4, it can be seen that a reduction in the total gas flow rate results in lower B concentration, for a given BCl_3 partial pressure. Although not shown here, this same behavior was also observed for Si_2H_6-source deposition, suggesting that Si precursor gas depletion is not the cause. (The higher RSC of Si_2H_6 means it should deplete faster than SiH_4, which, if depletion dominates, would suggest the opposite of what was observed.) Further supporting this case, in [12], it was shown that the RSC has an inverse relationship to gas flow, *i.e.* as the flow rate increases, RSC is reduced. For a constant pressure, increased gas flow results in increased gas velocity, which means that the amount of time a given gas molecule has to adsorb to the substrate surface is reduced. For the same BCl_3 partial pressure, then, a reduction in the total gas flow increases the amount of Si adsorbing to the surface, which reduces the B:Si ratio and therefore B concentration.

The dependence of Ge content on GeH_4 partial pressure for n- and p-type films is shown in Fig. 7.5. For p-type films, increased B concentration results in slightly reduced Ge content due to B enhancement of SiH_4 adsorption, as mentioned previously. As noted in [13], B adatoms increase the potential at the deposition surface, which has two effects: H more readily desorbs

from the surface, and SiH_4 molecules have some electrostatic attraction to the surface. These two effects combine to result in a nonlinear effect of B on SiH_4 adsorption, which explains the tapering off of the B concentration curves in Fig. 7.4.

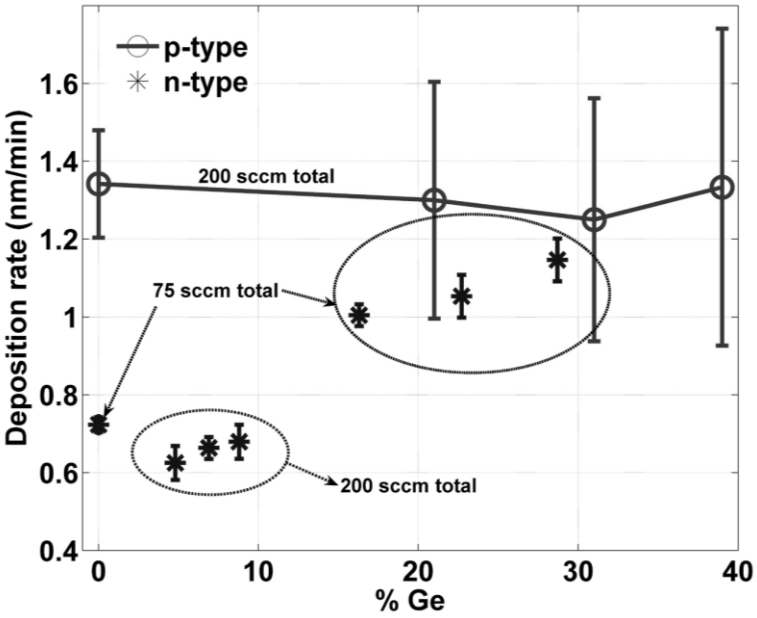

Fig. 7.6. Deposition rate vs. Ge content for n-type and p-type films. SiH_4 was used as the Si source gas for the p-type films; Si_2H_6 was used as the Si source gas for the n-type films. The shift between the 75 sccm (recipes NS1, Table I and NSG4, Table II) and 200 sccm (recipe NSG3, Table II) total gas flow cases for n-type films arises from the RSC dependence on gas flow rate, while the variation in p-type data (recipes PS3, Table I and PSG1, Table II) is due to the dependence of deposition rate on BCl_3 partial pressure.

For n-type films, in most cases, P concentration has little or no effect on Ge content, as Fig. 7.2(b) shows. Ge incorporation into the film seems to be primarily a function of GeH_4 partial pressure, which is shown in Figs. 7.2(b) and 7.5. Since P adatoms are n-type, their result is a decrease in the potential at the deposition surface, suggesting a reduction in H desorption as well as less electrostatic attraction of incoming Si_2H_6 molecules. Since the RSC of Si_2H_6 is much higher than that of SiH_4, however, the deposition rate of the n-type film should not change by much if at all, so long as any reduction in H desorption does not reduce the number of available adsorption sites to below what would normally be filled by Si_2H_6 if there were little or no P adatoms at the surface**.** This is supported by the constant deposition rate (0.7-0.75 nm/min, shown in Fig. 7.6) measured for n-type Si films deposited with various PH_3 partial pressures. Only in the highest P concentration ($6.4x10^{20}$ cm^{-3}) and Ge content (~30%) case presented, as Fig. 7.2(b) shows, does P incorporation seem to have a non-negligible effect on Ge content, which increases due to reduced Si_2H_6 adsorption. Thus, it would seem that the effect of P in changing the surface potential does not dominate over the high RSC of Si_2H_6 unless extremely high P concentrations are present in the deposited film. However, since the same effect can also be explained as the onset of P poisoning due to P-P dimerization, which may result in the preferential adsorption of GeH_4 over Si_2H_6, it is difficult to conclude with the current data whether the slight increase in Ge content for the highest P concentration case is due primarily to a reduction in surface potential.

From Fig. 7.5 it can also be seen that, for similar doping levels, Ge incorporation is much lower for n-type films, for a given GeH_4 partial pressure. To counteract this effect, higher GeH_4 partial pressures can be used, either by increasing the GeH_4 flow rate and/or reducing the total gas flow rate. The former approach increases the cost of deposition, while the latter approach may change the Ge content dependence on GeH_4 partial pressure due to a flow rate dependence of GeH_4 depletion within the tube [14] (i.e. RSC dependence on flow rate). This depletion effect is evident in Fig. 7.5, where higher Ge content in the n-type films was achieved for reduced total gas flow rate. Although higher Ge content can be achieved for n-type films this way, the required GeH_4 partial pressure is substantial and the sensitivity of Ge content on GeH_4 partial pressure is reduced at lower total gas flow rates (noting the log scale of the x-axis).

Since the deposition rate for n-type $Si_{1-x}Ge_x$ films is lower than that for p-type films of the same Ge content (Fig. 7.6), the reduction in Ge incorporation in n-type films is due to competition between PH_3 and GeH_4 for adsorption sites. (This is further supported by the smaller slope in Fig. 7.5 for n-type films for the same GeH_4 partial pressure range). P concentration has little effect on film deposition rate over the measured range of P and Ge concentrations; the variation in measured n-type film deposition rate in Fig. 7.6 is due primarily to the limited depth resolution of SIMS. For n-type films, then, the deposition rate depends on the GeH_4 partial pressure. This is clearly shown in Fig. 7.6, where the same linear dependence of the n-type deposition rate on Ge content is maintained between low (75 sccm) and high (200 sccm) total gas flow rates. That this linear dependence is maintained, and that the deposition rate curve shifts downward with a higher total flow rate, shows that the RSC dependence on flow rate dominates over the GeH_4 depletion dependence on flow rate in determining the deposition rate, and that the deposition is therefore not source-limited. (It is noted that preferential sputtering of different species during SIMS analysis results in some inaccuracy in film thicknesses determined by SIMS analysis alone; however, the trends observed in Fig. 7.6 should remain). In contrast, for p-type films, the deposition rate shows statistically insignificant dependence on Ge content, but significant dependence on BCl_3 partial pressure, which is why the range of measured deposition rates in Fig. 7.6 is much larger for p-type films.

7.3 SPER of LPCVD $Si_{1-x}Ge_x$ Films

Crystalline $Si_{1-x}Ge_x$ layers are formed by epitaxial growth, which tends to be expensive due to the need for ultra-high vacuum (UHV) conditions during the deposition process [8], [16]. A possible alternative is to leverage the lower cost of LPCVD to deposit amorphous or polycrystalline $Si_{1-x}Ge_x$ on a crystalline substrate, which is then used as a seed layer for solid phase epitaxial re-growth (SPER) [6], [17]-[18]. In the presented work, a thin (target thickness = 10-15 nm) LPCVD in-situ doped $Si_{1-x}Ge_x$ layer is deposited on bulk (100) Si substrates using the process outlined in Section 7.2, followed by LPCVD Si capping layer (target thickness = 25 nm). A Si implant ($5x10^{15}$ cm^{-2}, 50 keV, 0 º tilt) was then performed to completely amorphize the $Si_{1-x}Ge_x$/Si film stack as well as the substrate-to-$Si_{1-x}Ge_x$ interface. A subsequent furnace anneal in N_2 at 550 ºC for 12 hr was then performed to achieve SPER.

Fig. 7.7(a) shows an example TEM cross-section of a post-annealed $Si_{0.7}Ge_{0.3}$ sample, while Fig. 7.8(a) shows Ge SIMS data for $Si_{0.5}Ge_{0.5}$ samples after the same anneal condition. The annealed sample has indeed recrystallized, as evidenced by the (111) crystalline defects in the $Si/Si_{1-x}Ge_x$ stack. Higher SPER anneal temperatures reduced the presence post-anneal crystalline

defects (Fig. 7.7(b)), but at the cost of significant Ge diffusion (Fig. 7.8(b)). In fact, even for the 550 ºC/12 hr anneal, Ge diffusion reduces the peak from 45-55 % (pre-anneal) down to 15-20% (post-anneal). This significant Ge loss from the as-deposited $Si_{1-x}Ge_x$ layer appears to be defect-assisted, since Ge diffusion in amorphous Si (or poly-Si with small grain boundaries) is several orders of magnitude higher than in crystalline Si [19], suggesting that an amorphization implant and subsequent SPER anneal for LPCVD $Si_{1-x}Ge_x$ layers is fundamentally unsuitable for inexpensive formation of crystalline $Si_{1-x}Ge_x$ layers with moderate-to-high Ge content.

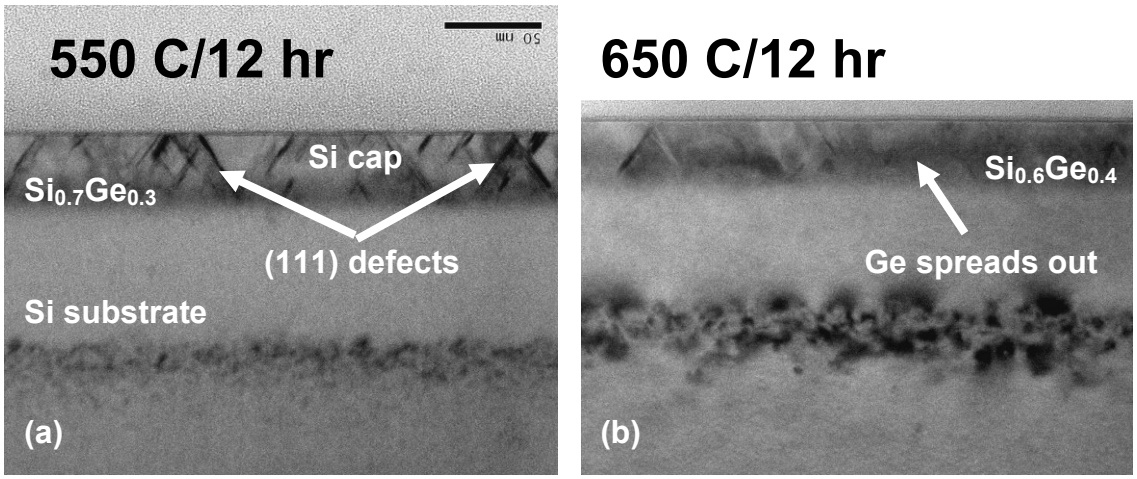

Fig. 7.7. TEM cross-sections of $Si/Si_{1-x}Ge_x$ samples with Si amorphization implant and subsequent SPER furnace anneal at (a) 550 ºC/12 hr and (b) 650 ºC/12 hr.

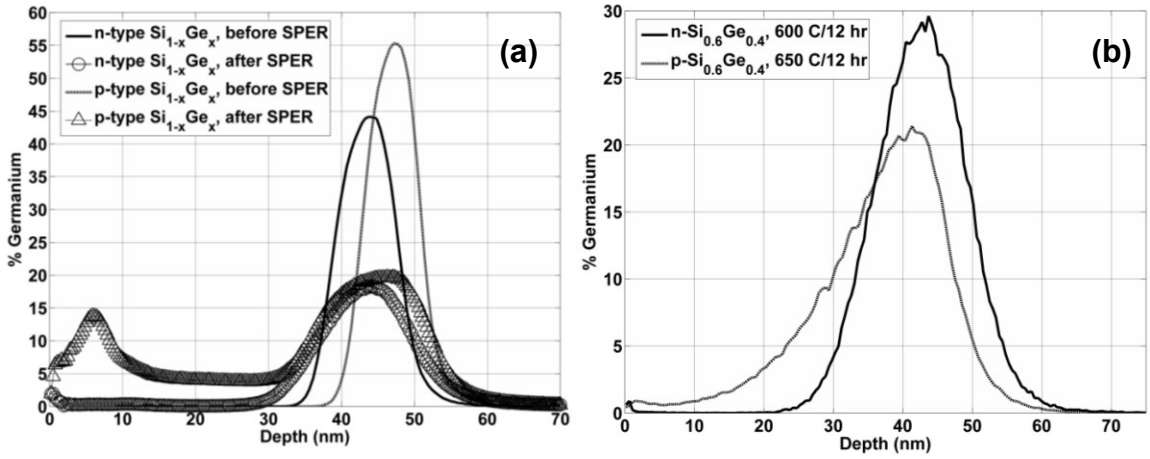

Fig. 7.8. Ge SIMS data for (a) n- and p-type $Si_{1-x}Ge_x$ annealed at 550 ºC/12 hr and (b) n-type $Si_{0.6}Ge_{0.4}$ annealed at 600 ºC/12 hr and p-type $Si_{0.6}Ge_{0.4}$ annealed at 650 ºC/12 hr.

The crystalline defects shown in Fig. 7.7 likely arise from O contamination within the as-deposited $Si/Si_{1-x}Ge_x$ layers (Fig. 7.9), which form SiO_x precipitates [20] during the SPER anneal. As the SIMS data in Fig. 7.9 shows for a multi-layer film stack, the residual O concentration in the undoped and P-doped Si layers is ~2-3x10^{19} cm^{-3} and increases to ~6-7x10^{19} cm^{-3} for B-doped Si (due to either the increased surface potential in B-doped Si or some interaction between B and O). Most notably, the O concentration spikes during $Si_{1-x}Ge_x$

deposition to over $1x10^{21}$ cm^{-3}. This is due to Ge readily oxidizing before or during Ge adsorption at the deposition front. The source of this high background O is unclear, but at least some of it is due to an impure N_2 gas line running to the LPCVD chamber. This is evidenced in Fig. 7.10, where a $Si_{1-x}Ge_x$ multi-layer film stack was deposited at multiple temperatures. The temperature was ramped and stabilized with the wafers in the N_2-purged LPCVD chamber. The large O spikes in Fig. 7.10 at depths of 50 and 125 nm indicate the presence of O within the N_2 gas line. If this parasitic O can be eliminated, presumably by replacing the N_2 gas line or otherwise finding and eliminating the leakage point/s within the line, then low temperature SPER of LPCVD $Si_{1-x}Ge_x$, without crystalline defects, should be possible.

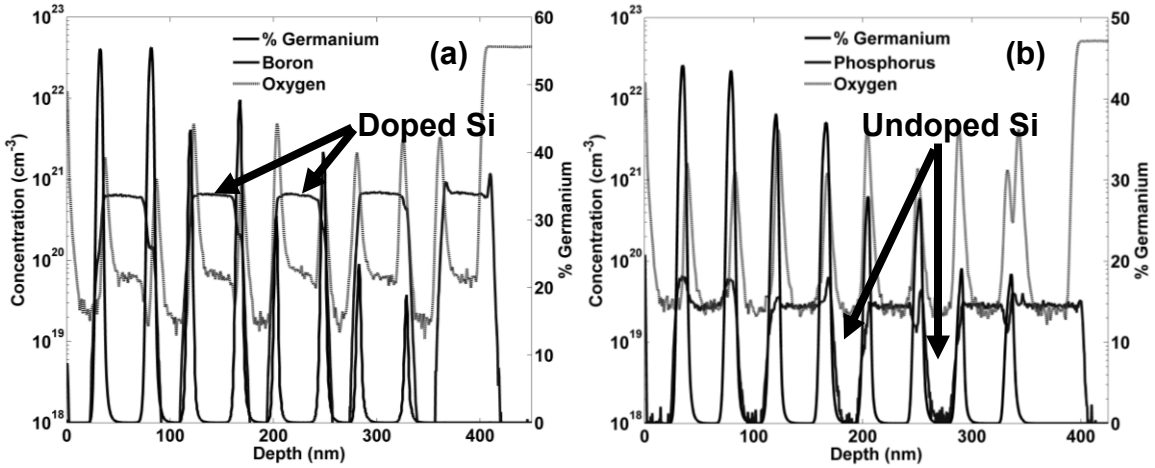

Fig. 7.9. SIMS data for (a) p-type and (b) n-type multi-layer LPCVD $Si/Si_{1-x}Ge_x$ film stacks deposited on SiO_2, with Ge content varying from 0 % to 50 %.

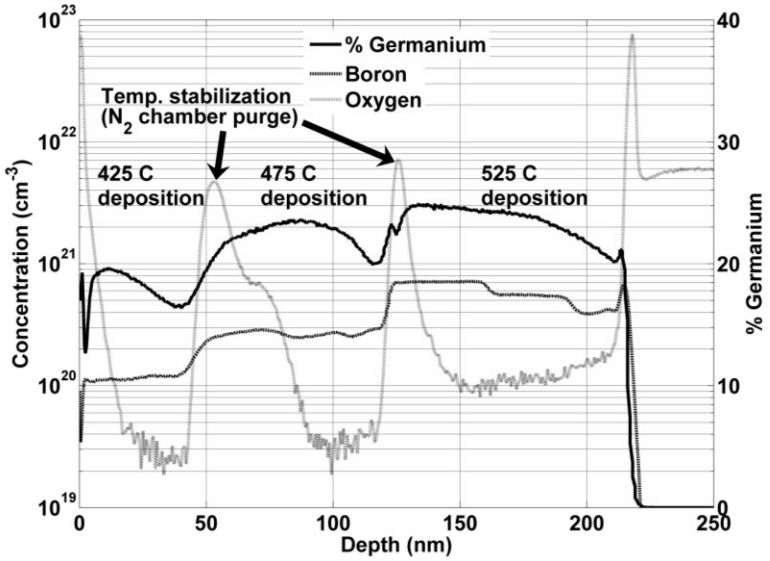

Fig. 7.10. SIMS data for LPCVD p-type $Si_{1-x}Ge_x$ layers deposited at 525 ºC, 475 ºC, and 425 ºC. The O spikes correspond to temperature stabilization steps, during which the LPCVD chamber is flooded with N_2.

7.4 Ge Melt Processing

An alternative approach to creating crystalline $Si_{1-x}Ge_x$ layers out of amorphous or polycrystalline $Si_{1-x}Ge_x$ layers is liquid phase epitaxy (LPE) of Ge, also known as Ge melt [21]-[26]. In this process, Ge (or $Si_{1-x}Ge_x$ with high Ge content, to reduce the melting temperature) is deposited onto a crystalline Si or $Si_{1-x}Ge_x$ substrate and subsequently heated to or above its melting temperature (938 ºC for pure Ge), but below the melting temperature of the substrate. As the Ge liquefies, some of the underlying Si dissolves into the melt. The Si dissolution into the melt increases until what is now a liquid $Si_{1-x}Ge_x$ melt can no longer dissolve more Si while remaining in a liquid state (Fig. 7.11). As the melt cools, Si atoms are rejected from the melt at the substrate-to-liquid-Ge interface, resulting in solidification of the melt from this interface out to the top of the as-deposited Ge layer, with the Si substrate as the seed layer. Thus, the portion of the Si substrate that was dissolved into the melt is replaced with a graded, crystalline $Si_{1-x}Ge_x$ layer, on top of which lies crystalline Ge.

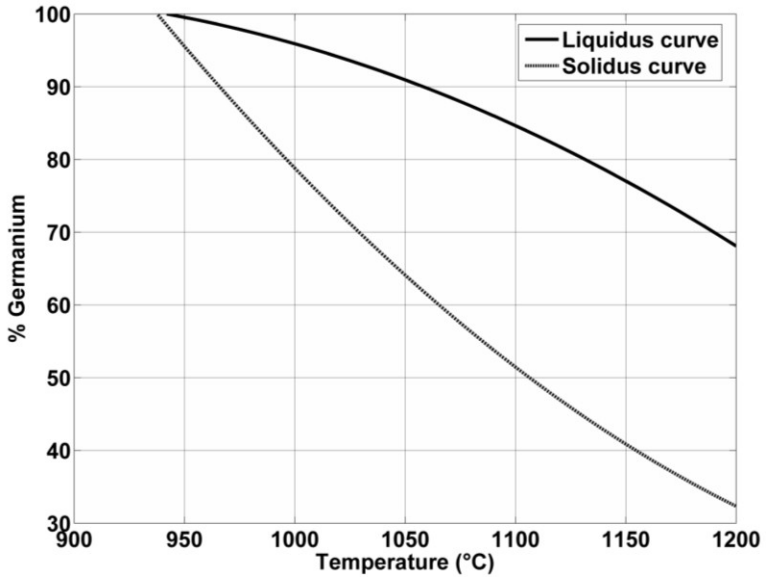

Fig. 7.11. Liquidus-solidus curves for the Si-Ge system.

Over the range of 938 ºC to 1200 ºC in Fig. 7.11, second-order polynomial fits can be attained for the liquidus and solidus curves. These are, respectively,

$$\%Ge_{liquid} = -83.1 + 0.444T - (2.65x10^{-4})T^2 \tag{7.1}$$

$$\%Ge_{solid} = 803.1 - 1.1343T + (4.1x10^{-4})T^2 \tag{7.2}$$

where T is the temperature in ºC. Thus, for an example of 1050 ºC, the melt will contain 91 % Ge and 9 % Si. This means that, for every 100 nm of Ge deposited, 9/0.91 = 9.89 nm of Si will dissolve into the melt. If the melt is cooled down in a quasi-steady state fashion (*i.e.*, very low ramp rate), then the Ge content of the graded $Si_{1-x}Ge_x$ region will follow the solidus curve. So, for the 1050 ºC example, the $Si_{1-x}Ge_x$-to-substrate interface (*i.e.*, the bottom of the 9.89 nm graded $Si_{1-x}Ge_x$ region) will contain $Si_{1-x}Ge_x$ with a Ge content of 64 %. As the solidification

progresses, the Ge content in the graded $Si_{1-x}Ge_x$ layer increases until there is no remaining Si in the solidifying melt, leading to pure, crystalline Ge. Contrarily, if the melt is cooled down rapidly, then there is not sufficient time for the Si in the melt to diffuse to the liquid/solid interface and be rejected. This will lead to the entire melt solidifying as-is, with a uniform Si concentration.

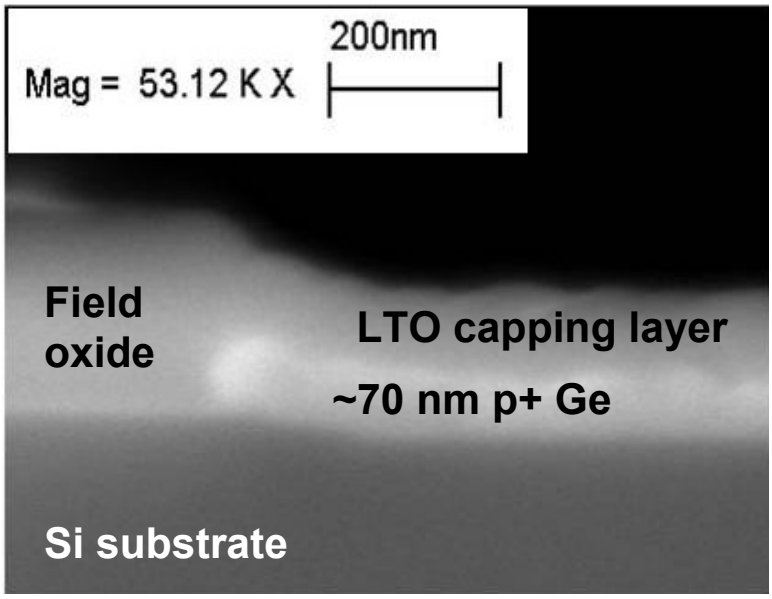

Fig. 7.12. SEM cross-section of as-deposited LPCVD p+ Ge.

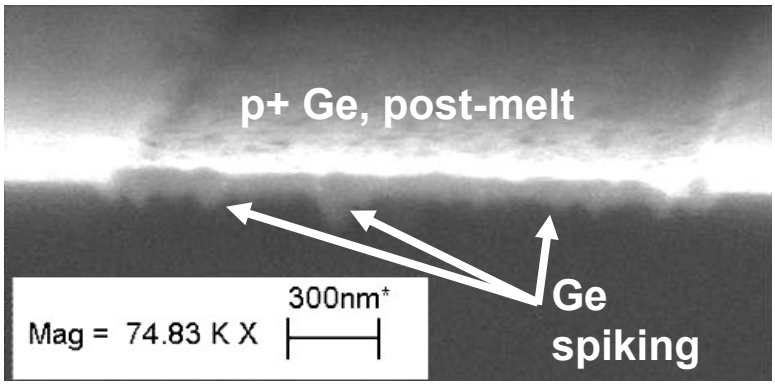

Fig. 7.13. SEM cross-section of LPCVD p+ Ge after 1000 ºC, 10 s RTA melt in N_2.

Fig. 7.12 shows a SEM cross-section of as-deposited LPCVD p+ Ge (425 ºC, 400 mT, 15 sccm GeH_4, 35 sccm BCL_3, 10 min. deposition), which is selectively deposited on a Si substrate between oxide trenches and then capped with low temperature oxide (LTO). The Ge-Si interface is smooth and the Ge surface shows relatively little roughness. Fig. 7.13 shows a SEM cross-section of the same film after a 1000 ºC, 10 s RTA melt in N_2 (and with the LTO capping layer removed). The post-melt sample shows significant amounts of Ge spiking into the substrate, on the order of 100 nm, as well as a rough, pitted surface, which correlates to the presence of Ge spikes [27].

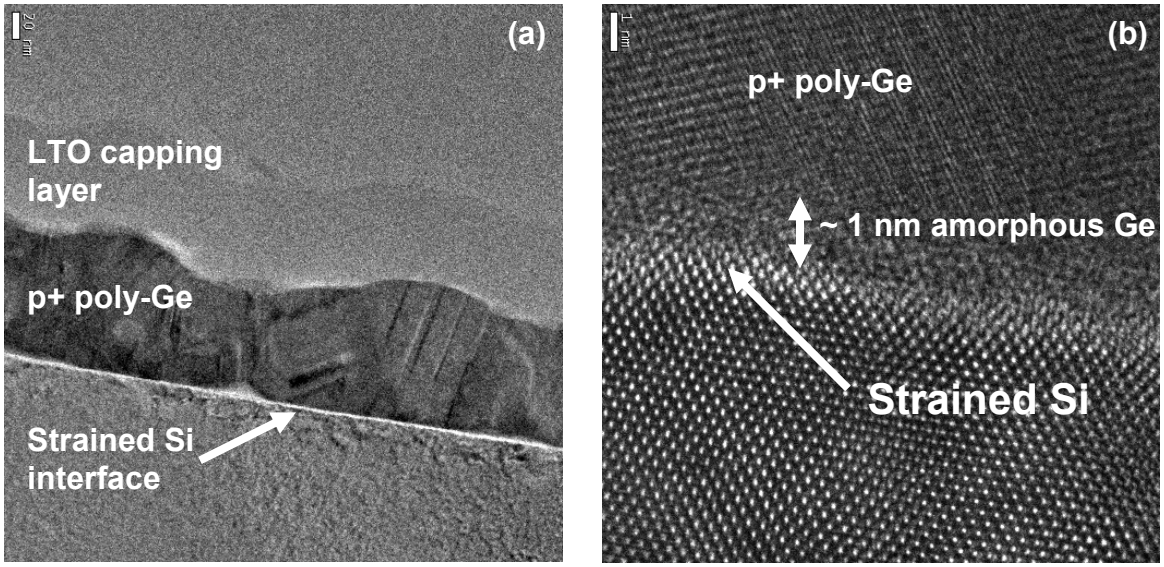

Fig. 7.14. (a) TEM cross-section of as-deposited LPCVD Ge film (same recipe as the sample from Fig. 7.12) and (b) high resolution TEM cross-section of the Ge-Si interface.

The source of this Ge spiking effect is traceable to significant and non-uniform strain at the Ge-Si interface for the as-deposited Ge film. As Fig. 7.14 shows, there is significant and non-uniform strain at the Ge-Si interface, despite the ~1 nm of amorphous Ge between the polycrystalline Ge layer and the Si substrate. This strain relaxes during the non-steady-state RTA ramp-up phase in the form of crystalline defects/stacking faults and gives rise to enhanced Ge diffusion along the stacking fault boundaries [19]. Assuming to a first order that the SEM contrast cuts off at 70 % Ge in the $Si_{1-x}Ge_x$ layer, the Ge spiking depth can be calculated using Equations 7.3 and 7.4, where t_{melt} is the ramp time to the Ge melting temperature of 938 ºC, R is the RTA ramp rate (50 ºC/s), and the initial temperature is 400 ºC or 673 K (all other terms have their usual meaning). Using the 40 nm grain size case from [19], the Ge spiking depth is calculated to be 145 nm after the ramp-up phase alone, which agrees well with the extent of Ge spiking in Fig. 7.13.

$$(Dt)_{ramp} = \int_0^{tmelt} tD_0 \exp\left(\frac{-E_A}{k(673+Rt)}\right)dt \tag{7.3}$$

$$X_{j,70\%} = 2\sqrt{(Dt)_{ramp}\ln\left(\frac{1}{0.7}\right)} \tag{7.4}$$

There are a few approaches which can lead to reduced Ge spiking during the Ge melt anneal. One approach is to use RTA melt only for devices with ultra-thin body regions [25], which are more flexible than bulk substrates and can therefore withstand higher strain. Another approach is to use an amorphous Si interlayer between the Si substrate and the LPCVD Ge layer [27]. This can be achieved by a pre-amorphization implant and/or depositing amorphous Si before the Ge deposition step. The amorphous Si interlayer decouples the strain between the Si substrate and the Ge layer, and this decoupling of strain is expected to grow and eventually saturate as the

thickness of the amorphous interlayer increases. Yet another approach is to perform a furnace melt (Fig. 7.15) anneal rather than RTA. The slow ramp rate of the furnace anneal creates a quasi-steady-state condition which results in uniform, solid-state Ge diffusion into the substrate before the Ge layer melts. This Ge diffusion creates a graded $Si_{1-x}Ge_x$ layer, which continues to diffuse during the melt anneal. If the graded $Si_{1-x}Ge_x$ layer is thick enough by the time the cool-down phase begins, all of the strain-induced crystalline defects will be contained within the graded $Si_{1-x}Ge_x$ "buffer" layer. This should lead to a defect-free post-melt Ge layer, since now the critical thickness of the Ge layer is defined by the Ge content at the top of the graded $Si_{1-x}Ge_x$ buffer layer. This is shown in Fig. 7.15, with the only exception being that the top layer is $Si_{0.15}Ge_{0.85}$ (determined from energy-filtered TEM or EFTEM, as in Fig. 7.16) rather than pure Ge. This is because the heater elements in the furnace are shut off for the cool-down phase, meaning the initial temperature drop from 1000 ºC is somewhat rapid. Thus, the resulting Ge content in the post-melt sample falls somewhere between the liquidus and solidus curves from Fig. 7.11.

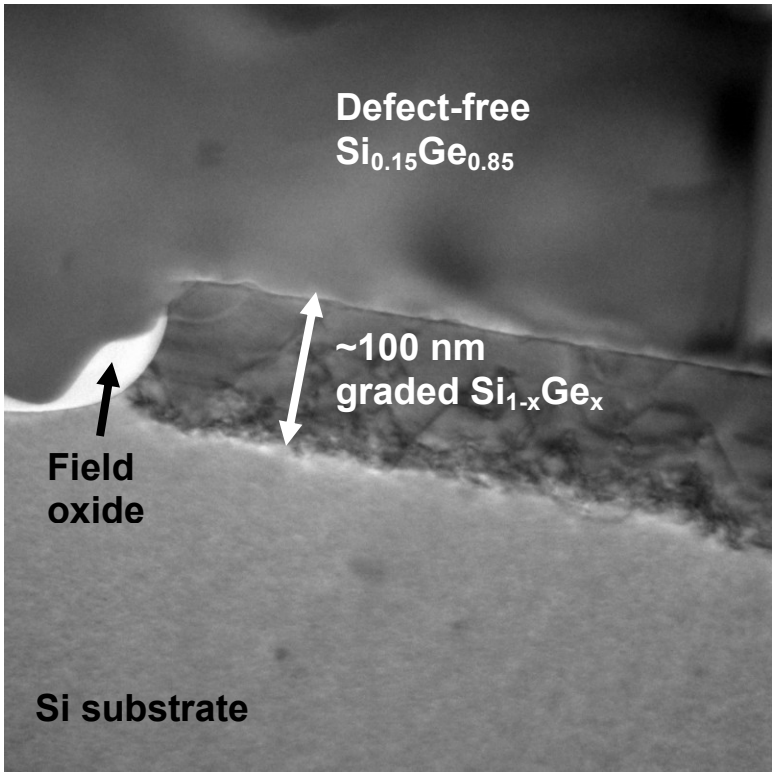

Fig. 7.15. TEM cross-section of furnace-annealed Ge-on-Si at 1000 ºC for 1 hr in N_2.

The size of the graded $Si_{1-x}Ge_x$ region in Fig. 7.15 is much larger than what is expected from classical Ge melt theory. The as-deposited Ge thickness was ~400 nm, meaning the thickness of the dissolved Si should be ~4 nm at 1000 ºC, which is much smaller than the ~100 nm graded $Si_{1-x}Ge_x$ layer. This supports the notion that significant Ge diffusion took place during the furnace ramp-up phase. Furthermore, the shape of the Ge profile determined from EFTEM (Fig. 7.16) does not follow a simple Gaussian or erfc relationship, suggesting that this pre-melt Ge diffusion is either stress-enhanced and/or Ge-dependent, being higher for larger Ge concentrations.

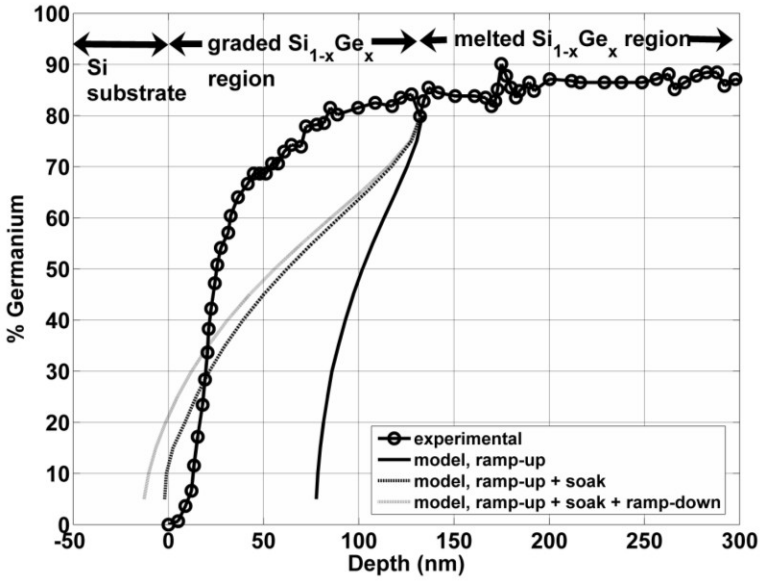

Fig. 7.16. Model vs. data for the furnace annealed sample in Fig. 7.15.

To model the effect of Ge-dependent Ge diffusion, D_0 and E_A as a function of Ge content are taken from [28] for relaxed $Si_{1-x}Ge_x$ (first order approximation). Only values for 0-50% Ge are reported in [28], but diffusivity D at a given temperature versus Ge content shows a log-linear relationship. This relationship is extrapolated to 100 % Ge to estimate D for higher Ge contents and is expressed as Equation (7.5), which permits Ge diffusion modeling in steady-state at 1000 ºC.

$$D_{Ge,1000C} = 1.234x10^{-16} * \exp(0.1029 * \%Ge) \tag{7.5}$$

This D vs. % Ge relationship is repeated over a range of temperatures to obtain a single, best-fit relationship for D as a function of Ge content and temperature, which permits modeling of Ge diffusion during thermal ramping. This relationship for incremental Dt product (in units of cm^2) is expressed as

$$Dt_{Ge,incr} = 1.69x10^{-128}(700 + Rt)^{37.312} * \exp\left(\%Ge * 0.71358 * \exp\left[-1.93x10^{-3}(700 + Rt)\right]\right) \tag{7.6}$$

where t is the time in min. This expression is then integrated over the ramp time to obtain a cumulative Dt product for a given Ge content, which is then repeated over a range of Ge contents from 0 % to 100 %. This results in a log-linear relationship for Dt product for a given ramp time and rate. For the example used here, a 5 ºC/min ramp rate is assumed from 700 ºC to 1000 ºC (*i.e.*, a 1 hr ramp), resulting in an expression for cumulative Dt product vs. Ge content as

$$(Dt)_{ramp} = 4.3602x10^{-14} * \exp(0.11097 * \%Ge) \tag{7.7}$$

Ge diffusion is modeled using erfc diffusion, since the Ge diffuses from a semi-infinite source

of Ge. To predict the Ge profile, a user-defined value for Ge content is inserted into Equations (7.5) and (7.7). Then, a erfc profile is modeled and the depth at which the modeled profile reaches the user-defined % Ge value is treated as the junction depth for that particular % Ge value. This process is repeated over a range of % Ge values to obtain the final Ge profile. The modeled Ge profiles are compared to experimental data in Fig. 7.16. Although the model and data do not match perfectly, the trend is the same, showing a pleateau at higher Ge contents which then drops off sharply at low Ge contents. The model-to-data comparison suggests that the Ge diffusivity at higher Ge contents is higher than what the model predicts, perhaps due to compressive strain in the direction of diffusion [28]. It is also possible that at least some of the difference may be due to Ge enrichment of the graded $Si_{1-x}Ge_x$ layer during the melt process [24], because Si dissolution into the melt must take place over a larger depth in consideration of the lower Si content in the graded $Si_{1-x}Ge_x$ region relative to a bulk Si substrate. Since the graded $Si_{1-x}Ge_x$ layer has ~70-85% Ge, the "consumed" depth of the substrate region due to Si dissolution increases to ~18 nm at 1000 ºC (compared to 4 nm for a pure Si substrate). Regardless, it is clear that the graded $Si_{1-x}Ge_x$ buffer layer forms primarily as the result of solid-state Ge diffusion during the slow furnace ramp-up and soak time, and that this buffer layer minimizes post-melt defects in the top $Si_{1-x}Ge_x$ layer.

7.5 Summary

Low temperature LPCVD of in-situ doped n- and p-type $Si_{1-x}Ge_x$, as well as crystallization schemes for as-deposited $Si_{1-x}Ge_x$ and Ge films, has been investigated in this study. The $Si_{1-x}Ge_x$ deposition rate and dopant incorporation are purely attributable to surface reaction phenomena such as adsorption site formation, competition for adsorption sites, and RSC. For n-type $Si_{1-x}Ge_x$ deposition, the deposition rate is determined primarily by the GeH_4 partial pressure. Also, Si_2H_6 is the preferred Si source gas, since the higher RSC compared to SiH_4 pushes the P poisoning concentration out to higher P concentrations. For p-type $Si_{1-x}Ge_x$ deposition, the deposition rate is determined primarily by the BCl_3 partial pressure, and SiH_4 is the preferred Si source gas as it increases the stability of dopant incorporation. Additionally, dopant incorporation has different dependencies for n- and p-type $Si_{1-x}Ge_x$: it increases with Ge content for n-type films while decreasing with increasing Ge content for p-type films. Amorphization implants followed by SPER for these LPCVD $Si_{1-x}Ge_x$ films has been demonstrated, but a trade-off exists Ge diffusion within the amorphous film (driving the need for low temperature SPER) and parasitic O within the as-deposited film (driving the need for high temperature SPER). Ge melt is proposed as an alternative to SPER, to form crystalline $Si_{1-x}Ge_x$ with high Ge content. A furnace melt anneal is shown to be more effective than RTA, since stress-induced Ge spiking is reduced and the graded $Si_{1-x}Ge_x$ buffer layer formed by Ge diffusion during the slow furnace ramp minimizes post-melt stress between the top $Si_{1-x}Ge_x$ layer and the Si substrate.

7.6 References

[1] W.-C. Lee, Y.-C. King, T.-J. King, and C. Hu, "Investigation of poly-$Si_{1-x}Ge_x$ for dual gate CMOS technology," *IEEE Electron Device Letters*, vol. 19, no. 7, pp. 247-249, 1998.

[2] Y. V. Ponomarev, P. A. Stolk, C. Salm, J. Schmitz, P. H. Woerlee, "High-performance deep submicron CMOS technologies with polycrystalline-SiGe gates," *IEEE Transactions on Electron Devices*, vol. 47, no. 4, pp. 848-855, 2000.

[3] A. Muto, H. Ohji, T. Kawahara, T. Maeda, K. Torii, and H. Kitajima, "Improved performance of FETs with $HfAlO_x$ gate dielectric using optimized poly-SiGe gate electrodes," *Extended Abstracts of International Workshop on Gate Insulator*, pp. 64-68, 2003.

[4] A. E. Franke, J. M. Heck, T.-J. King, and R. T. Howe, "Polycrystalline silicon-germanium films for integrated microsystems," *Journal of Microelectromechanical Systems*, Vol. 12, pp. 160-171, 2002.

[5] H. Takeuchi, A. Wung, X. Sun, R. T. Howe, and T.-J. King, "Thermal budget limits of quarter-micron foundry CMOS for post-processing MEMS devices," *IEEE Transactions on Electron Devices*, Vol. 52, No. 9, pp. 2081-2086, 2005.

[6] Y. C. Yeo, V. Subramanian, J. Kedzierski, P. Xuan, T.-J. King, J. Bokor, and C. Hu, "Nanoscale ultra-thin-body silicon-on-insulator P-MOSFET with a SiGe/Si heterostructure channel," *IEEE Electron Device Letters*, vol. 21, no. 4, pp. 161-163, 2000.

[7] S.-M. Jang, K. Liao, and R. Reif, "Phosphorus doping of epitaxial Si and $Si_{1-x}Ge_x$ at very low pressure," *Applied Physics Letters*, vol. 63, no. 12, pp. 1675-1677, 1993.

[8] M. Racanelli and D. W. Greve, "*In situ* doping of Si and $Si_{1-x}Ge_x$ in ultrahigh vacuum chemical vapor deposition," *Journal of Vacuum Science and Technology B*, vol. 9, no. 4, pp. 2017-2021, 1991.

[10] C. W. Low, M. L. Wasilik, H. Takeuchi, T.-J. King, and R. T. Howe, "In-situ doped poly-SiGe LPCVD process using BCl_3 for post-CMOS integration of MEMS devices," *SiGe Materials, Processing, and Device Symposium*, Electrochemical Society, Honolulu, Hawaii, October 3-8, 2004.

[11] Y.-C. Jeon, T.-J. King, and R. T. Howe, "Properties of phosphorus-doped poly-SiGe films for microelectromechanical system applications," *Journal of the Electrochemical Society*, vol. 150, no. 1, pp. H1-H6, 2003.

[12] B. S. Meyerson and M. L. Yu, "Phosphorus-doped polycrystalline silicon via LPCVD, II. Surface interactions of the silane/phosphine/silicon System," *Journal of the Electrochemical Society*, vol. 131, no. 10, pp. 2366-2368, 1984.

[13] R. J. Buss, P. Ho, W. G. Breiland, and M. E. Coltrin, "Reactive sticking coefficients for silane and disilane on polycrystalline silicon," *Journal of Applied Physics*, vol. 63, no. 8, pp. 2808-2819, 1988.

[14] C.-A. Chang, "On the enhancement of silicon chemical vapor deposition rates at low temperatures," *Journal of the Electrochemical Society*, vol. 123, no. 8, pp. 1245-1247, 1976.

[15] A. Kovalgin, J. Holleman, "Low-Temperature LPCVD of Polycrystalline Ge_xSi_{1-x} Films with High Germanium Content," *Journal of the Electrochemical Society*, vol. 153, no. 5, pp. G363-G371, 2006.

[16] B. S. Meyerson, "UHV/CVD Growth of Si and Si:Ge Alloys: Chemistry, Physics, and Device Applications," *Proc. IEEE*, vol. 80, no. 10, pp. 1592-1608, Oct. 1992.

[17] S. Yamaguchi, N. Sugii, S. K. Park, "Solid-phase crystallization of $Si_{1-x}Ge_x$ alloy layers," *J. Appl. Phys.*, vol. 89, no. 4, pp. 2091-2095, Feb. 2001.

[18] M. Y. Tsai, B. G. Streetman, "Recrystallization of implanted amorphous silicon layers. I. Electrical properties of silicon implanted with BF_2^+ or $Si^+ + B^{+a)}$," *J. Appl. Phys.*, vol. 50, no. 1, pp. 183-187, Jan. 1979.

[19] A. Portavoce, G. Chai, L. Chow, J. Bernardini, "Nanometric size effect on Ge diffusion in polycrystalline Si," *J. Appl. Phys.*, vol. 104, pp. 104910, 2008.

[20] S. Wolf, R. N. Tauber, "Silicon Processing for the VLSI Era, Volume 1 – Process Technology, 2nd ed.," *Lattice Press*, 2000, pp. 49-53.

[21] Y. Liu, M. D. Deal, J. D. Plummer, "High-quality single-crystal Ge on insulator by liquid-phase epitaxy on Si substrates," *Appl. Phys. Lett.*, vol. 84, no. 14, pp. 2563-2565, Apr. 2004.

[22] S. Balakumar, M. M. Roy, B. Ramamurthy, C. H. Tung, G. Fei, S. Tripathy, C. Dongzhi, R. Kumar, N. Balasubramanian, D. L. Kwong, "Fabrication Aspects of Germanium on Insulator from Sputtered Ge on Si-Substrates," *Electrochemical and Solid-State Letters*, vol. 9, no. 5, pp. G158-G160, 2006.

[23] J. Zhao, A. C. Seabaugh, T. H. Kosel, "Rapid Melt Growth of Germanium Tunnel Junctions," *J. Electrochem. Soc.*, vol. 154, no. 6, pp. H536-H539, 2007.

[24] F. Lu, H.-S. Wong, K.-W. Ang, M. Zhu, X. Wang, D. M.-Y. Lai, P.-C. Lim, B. L. H. Tan, S. Tripathy, S.-A. Oh, G. S. Samudra, N. Balasubramanian, Y.-C. Yeo, "A New Source/Drain Germanium-Enrichment Process Comprising Ge Deposition and Laser-Induced Local Melting and Recrystallization for P-FET Performance Enhancement," *Symp. VLSI. Tech. Dig.*, pp. 26-27, 2008.

[25] T.-Y. Liow, K.-M. Tan, R. T. P. Lee, M. Zhu, B. L.-H. Tan, G. S. Samudra, N. Balasubramanian, Y.-C. Yeo, "5 nm Gate Length Nanowire-FETs and Planar UTB-FETs with Pure Germanium Source/Drain Stressors and Laser-Free Melt-Enhanced Dopant (MeltED) Diffusion and Activation Technique," *Symp. VLSI. Tech. Dig.*, pp. 36-37, 2008.

[26] N. Sugii, S. Yamaguchi, K. Washio, "SiGe-on-insulator substrate fabricated by melt solidification for a strained-silicon complementary metal-oxide-semiconductor," *J. Vac. Sci. Technol. B*, vol. 20, no. 5, pp. 1891-1896, Sep/Oct 2002.

[27] B. Ho., R. Vega, T.-J. King Liu, "Study of Germanium Epitaxial Recrystallization on Bulk-Si Substrates," *Proc. MRS Symp.* (to be published, 2010).

[28] N. R. Zanenberg, J. L. Hansen, J. Fage-Pedersen, A. N. Larsen, "Ge Self-Diffusion in Epitaxial $Si_{1-x}Ge_x$ Layers," *Phys. Rev. Lett.*, vol. 87, no. 12, pp. 125901, Sept. 2001.

Chapter 8

Conclusions

8.1 Summary

The requisite scaling in body thickness t_{body} with gate length L_G in thin-body MOSFETs, and concomitant increase in source/drain series resistance R_{SD}, motivates the transition from doped source/drain MOSFETs to metallic source/drain (MSD) MOSFETs; namely, dopant-segregated Schottky (DSS) source/drain MOSFETs. However, this transition in source/drain architecture is only appropriate for certain power/performance specifications. For low standby power (LSTP) design, DSS MOSFETs are not appropriate due to the dual ambipolar leakage mechanisms (SDE tunneling and band-to-band tunneling, or BTBT) which impose a small SDE design space that shrinks with L_G. A raised source/drain (RSD) structure is instead more appropriate for LSTP, because the contact interface is out-of-plane, thus eliminating SDE tunneling.

For high performance (HP) and low operating power (LOP) design, SDE doping can be very high due to the high I_{OFF} (HP design) or low V_{DD} (LOP) specification. These applications are more appropriate for DSS technology, where the leakage floor $I_{min} << I_{OFF}$. Still, second order effects such as silicide gating mandate careful sidewall spacer and SDE co-optimization in order to extract the most from the DSS architecture. Despite the fact that DSS thin-body MOSFETs have a small contact area at the silicide/silicon interface, the modeling work presented in this dissertation shows that contact resistance variation due to random dopant fluctuation (RDF) will not be significant, even at the end of the CMOS roadmap. This lends weight to the notion that NiSi will perform well enough as a single silicide material for CMOS (either in DSS form or conventional, doped source/drain form) to the end of the roadmap.

For the first time ever, it has been empirically demonstrated that the SDE junction depth $X_{j,SDE}$ in DSS MOSFETs can be modulated, for the same dopant species, silicide material, implant-to-silicide (ITS) conditions, and post-ITS anneal conditions, by exerting control over the silicide thermal stability. This was demonstrated with NiSi by performing a fluorine pre-silicidation ion implant (F-PSII), which passivates the NiSi/Si interface to reduce the average interface energy between the various NiSi grains and the adjacent Si. The result is a reduction in the amount of Ni rejection from the NiSi layer, thus sharpening the Ni diffusion profile into the Si, thus reducing the spatial vacancy distribution which accelerates dopant diffusion.

A new device structure has been proposed, the HTI Tri-Gate MOSFET, as an alternative to FinFETs for ultimate MOSFET scalability. The key feature of this device is that high-k dielectrics are used as the trench isolation or trench liner, to amplify gate fringing field coupling to the active region sidewalls. This relaxes the body doping requirements in bulk MOSFETs,

resulting in a device which can scale as well as a FinFET, without requiring the active width to be smaller than L_G as in a FinFET. Bulk LSTP scalability for the HTI Tri-Gate MOSFET has been demonstrated through TCAD to be competitive with FinFETs to the end of the CMOS roadmap.

Finally, $Si_{1-x}Ge_x$ process technology was explored and it was demonstrated that low temperature $Si_{1-x}Ge_x$ deposition for n- and p-type films is more complex than simply switching the dopant carrier gas. Different Si carrier gases are required for n- and p-type films, due to the difference in reactive sticking coefficient (RSC) between the dopant and Si carrier gases and the resulting effect on the stability of the deposition process. Forming crystalline $Si_{1-x}Ge_x$ by pre-amorphization implant of LPCVD $Si_{1-x}Ge_x$ and subsequent solid phase epitaxial regrowth (SPER) was shown to work, although high parasitic O levels in the as-deposited films resulted in defective films. Higher temperature SPER anneals reduce this defectivity, but at the cost of significant Ge diffusion which blurs the $Si_{1-x}Ge_x$/Si interface. Ge melt processing was also explored and was found to be a more promising approach to forming crystalline $Si_{1-x}Ge_x$ from LPCVD Ge, provided a slow ramp-up (*i.e.*, furnace anneal) process is used to minimize the Ge spiking effect.

8.2 Future Research Prospects

Further work on the topics presented in this dissertation fall into three topics: $Si_{1-x}Ge_x$ technology, DSS technology, and HTI technology. As regards $Si_{1-x}Ge_x$ technology, further experimentation on Ge melt is necessary. In particular, using Ge melt to form PMOS source/drain stressors is an interesting topic. It may be that, although Ge melt results in high Ge concentration in the resulting $Si_{1-x}Ge_x$ layer, a better approach would be Ge diffusion. In other words, performing a pre-amorphization implant into the PMOS source/drain regions, followed by Ge diffusion into and subsequent melting of these regions, may result in a higher quality source/drain stressor due to a reduced Ge spiking effect during RTA melting. Thus, the effect of an amorphous Si interlayer (in particular, the thickness of this interlayer) between the LPCVD polycrystalline Ge and the crystalline Si substrate on Ge spiking will aid in the optimization of Ge melt by RTA. The results of such an experiment could then be applied to PMOSFET source/drain stressor design, and it would be very interesting to determine whether this type of process can result in equivalent or higher PMOSFET performance than when epitaxial $Si_{1-x}Ge_x$ (which is a more expensive process) is used to form the source/drain stressors. Positive results of such an experiment may lead to lower cost high-performance PMOSFETs.

With regard to DSS technology, future ITS experiments should focus on the effect of F-PSII dose on $X_{j,SDE}$ for different dopant species and silicide materials. For example, how does As/NiSi ITS compare to As/PtSi ITS over a range of F-PSII doses? Is there a critical F-PSII dose beyond which $X_{j,SDE}$ is the same for As/NiSi and As/PtSi ITS? Although the experimental results in Chapter 6 suggest that such a critical dose may exist, further experimentation is necessary to build a database of ITS process conditions and corresponding $X_{j,SDE}$ and interface dopant concentration for a given post-ITS anneal temperature. Additionally, ITS for B and BF_2 should be compared directly, since the results in Chapter 6 suggest that B segregation at the NiSi/Si interface will require shorter anneals for B, since no F from a BF_2 implant would exist in the NiSi layer to retard B diffusion within the NiSi. Each of these experiments should be performed on bulk samples as well as very narrow samples, to determine whether there is a

dependence of $X_{j,SDE}$ on the width of the active region. If the active region is one silicide grain wide (*e.g.*, FinFET), and if dopant diffusion within the silicide takes place primarily between the grain boundaries, then for such a device the dopant diffusion will take place along the active sidewalls and may be "fast" or "slow," depending on whether some other material covers these sidewalls to interface the silicide region. Thus, the optimal post-ITS anneal time and/or temperature may change, meaning DSS FinFETs and DSS FDSOI planar MOSFETs may require different optimal ITS process conditions, even for the same L_G, sidewall spacer length, etc.

A more fundamental question about DSS technology which deserves attention is the effect of strain. Strain was ignored in this dissertation on purpose, in order to understand DSS vs. doped source/drain on a fundamental level. In a DSS structure, the only strain induced by the source/drain regions is that formed by the silicidation process. It is presently unclear whether the reduction in R_{SD} for DSS MOSFETs is enough to result in I_{ON} that is competitive with aggressively-scaled MOSFETs using $Si_{1-x}Ge_x$ or $Si_{1-y}C_y$ source/drain stressors which, by definition, cannot be DSS MOSFETs. Thus, there may be a fundamental trade-off here, between R_{SD} and channel strain. If so, then maybe DSS and $Si_{1-x}Ge_x$/$Si_{1-y}C_y$ stressors can be combined, whereby the SDE region in DSS MOSFETs consists of $Si_{1-x}Ge_x$ or $Si_{1-y}C_y$ with high Ge or C content, such that the benefits of DSS and of source/drain stressors can be realized in a single structure. This is a topic worthy of future modeling and experiment.

The purpose of the modeling study of HTI technology presented in Chapter 5 was simply to open peoples' minds to new and interesting ways to extend semi-planar MOSFET scalability. However, much work remains in this topic. On the modeling front, the effect of interface states at the HTI/active interface on BTBT leakage and gate control must be studied, as well as the effect of HTI on coupling strain to the active regions. Further along the lines of strain, the HTI vs. FinFET comparison in Chapter 5 assumes no strain; however, one of these devices will be better suited to strain and this may change the relative competitiveness of HTI vs. FinFETs, for better or worse. On the experimental front, there are a number of possible integration schemes for HTI and these must be explored to determine which scheme provides the highest performance for the least process complexity.

8.3 Conclusions

It was the goal of this dissertation to determine whether NiSi is suitable for single silicide CMOS to the end of the CMOS roadmap and to explore DSS and other MOSFET designs at and, in some cases beyond, their perceived scalability limits. Through modeling and experiment, it has been found that the main criticism of NiSi – being a near-midgap silicide with moderate electron and hole Schottky barrier heights (SBH), leading to high contact resistance – is not a limiting factor to its scalability. Dopant segregation is more than adequate in reducing the SBH to zero or near-zero values and so whatever ends up limiting NiSi scalability, it will not be contact resistance. Nor will it be RDF which, for dopant concentrations consistent with source/drain doping levels, results in a small spread in contact resistance even for contact areas at the end of the roadmap (20-30 nm^2). As far as MOSFET scalability is concerned, L_G is a misleading metric and instead the source-to-drain contact spacing L_{total} should be the point of focus. To this end, direct source-to-drain tunneling (DSDT) is significant for optimized designs at L_{total} ~ 10 nm. Below 10 nm, DSDT increases but never dominates over thermal leakage, meaning conventional short channel effects (SCE) will limit CMOS scaling. Regardless of the

device structure, be it FinFETs, planar FDSOI, semi-planar bulk, with DSS or doped source/drain regions, high-k dielectrics are critical to MOSFET scalability, but not necessarily as the gate dielectric. Properly implemented, high-k dielectrics function better as part of a dual high-k/low-k sidewall spacer and/or as the trench liner material. This is because, at these scales, the perimeter of the gate electrode is a substantial fraction of the overall device dimension and the fringing fields along the gate perimeter can and need to be leveraged to improve gate control (and device performance in general) in this regime. Without developing this high-k spacer and HTI technology further, CMOS scaling may very well end sooner than it needs to.