Design for Performance and Reliability in Advanced CMOS Structures

Fei Ding

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2020-41
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2020/EECS-2020-41.html

May 5, 2020

Copyright © 2020, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

Design for Performance and Reliability in Advanced CMOS Structures

by

Fei Ding

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 Junqiao Wu
Professor Ming Wu

Spring 2020

Design for Performance and Reliability in Advanced CMOS Structures

Copyright 2020

by

Fei Ding

Abstract

Design for Performance and Reliability in Advanced CMOS Structures

by

Fei Ding

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Tsu-Jae King Liu, Chair

In the past few decades, the central theme of the electronics industry is to increase the transistor density by reducing the transistor area, as required by the Moore's Law. The paradigm shift from the planar CMOS technology to the FinFET technology sustains this area scaling trend into sub-20nm era. The enhancement in the transistor electrostatics in the FinFET enables further scaling of the gate length and hence the contacted poly pitch (CPP). Meanwhile, the quest for area scaling also comes from the width (or fin pitch) and height dimensions. By reducing the fin pitch and increasing the fin height, the current density of the FinFET can be improved. Consequently, circuit designers can use fewer fins to meet the same current requirement and save area simultaneously, a scheme commonly referred to as "fin-depopulation." However, the aforementioned approaches start to show diminishing returns and meet excessive fabrication challenges. To further improve the current density and reduce the area, novel channel materials with high mobility (e.g., SiGe) and/or new structures with even better electrostatics (e.g., Inserted-oxide FinFET (iFinFET), Gate-All-Around FET, Nanosheet FET) are projected to be used in the future.

In the first part of the talk, the performance of a p-channel FinFET comprising a heterogeneous silicon (Si) and silicon-germanium ($Si_{0.9}Ge_{0.1}$) channel region is evaluated using 3-D TCAD simulations. It is shown that the hetero-channel design provides for larger current density while maintaining comparable electrostatic integrity as the conventional Si FinFET design due to the valence band (VB) offset between SiGe and Si.

Secondly, a scheme for controllably adjusting transistor drive strength in iFinFET technology is proposed, to enable cell ratio tuning for a minimally sized

six-transistor SRAM (6-T Static Random Access Memory) cell. It is demonstrated, via 3-D TCAD simulation, that this scheme can reduce the minimum cell operating voltage (V_{min}) and facilitate further cell area scaling.

Lastly, as the transistor area continues to shrink, self-heating effects of these small-geometry transistors have been of great concern as it limits the electrical performance and degrades the reliability of transistors. It is important to understand how self-heating may be for these new transistor structures as compared to FinFETs. The performance of advanced transistor structures (i.e., FinFET, Gate-All-Around FET, and Nanosheet FET) is simulated and compared under the constraints of the self-heating. An optimization guideline for nanosheet FETs is also proposed based on the study of various design parameters on the self-heating effects.

To my family

for their ceaseless support along the journey.

Contents

1 Introduction 1

1.1 IC Chip Area Scaling – The Central Theme 1

1.2 Transistor Design Techniques to Facilitate Further Area Scaling 6

1.3 Research Objectives and Thesis Overview 9

1.4 References 10

2 $Si/Si_{1-x}Ge_x/Si$ Hetero-Channel FinFET for Enhanced P-Channel Performance in Low-Power Applications 14

2.1 Introduction 14

2.2 Hetero-Channel ($Si/Si_{0.9}Ge_{0.1}/Si$) FinFET 15

2.3 Hetero-Channel FinFETs with Varying Ge Molefraction 22

2.4 Conclusion 24

2.5 Appendix: Process Simulation in Sentaurus Process 25

2.6 References 26

3 Cell Ratio Tuning for High-Density SRAM Voltage Scaling with Inserted-Oxide FinFETs (iFinFETs) 30

3.1 Introduction 30

3.2 Inserted-Oxide FinFET (iFinFET) 31

3.3 6-T SRAM High Density Cell Design 34

3.4 High Density Cell Ratio Tuning Using Doped-Nanowire iFinFET 41

3.5 Conclusion 54

3.6 Appendix: Doped-Nanowire iFinFET Fabrication Process Issues 54

3.7 References 58

4 A Comparison of Self-Heating Effects in Different Transistor Structures 63

4.1 Introduction 63

4.2 Simulation Methodology 64

4.3 Results and Discussion 67

4.4 Optimization of P-Channel NSF 78

4.5 Conclusion 94

4.6 Appendix: Validity of Using a 1-Fin FF/GAAF with AreaFactor=4 to Simulate 4-Fin FF/GAAF 96

4.7 References 98

5 Conclusion 101

5.1 Summary and Contributions of This Work 101

5.2 Future Directions 104

5.3 References 105

Acknowledgements

During the past four and a half years as a Ph.D. student, I am very fortunate to meet and work with a group of outstanding people. As I am concluding the entire course of my graduate study with this dissertation, I would like to use this opportunity to thank all the people that have helped me along the journey. Without them, this dissertation would not have been possible.

First and foremost, I am extremely grateful to my thesis advisor, Prof. Tsu-Jae King Liu, for her excellent mentorship and guidance throughout my graduate study. Her broad and deep knowledge in semiconductor device technology can always point us to the right direction when it comes to solving a difficult research problem. In addition, her enthusiasm in this field has always motivated us to explore new opportunities and take on new challenges. Outside research, she is also a role model for us. Her elegance, kind personality, and efficient use of time have always set an example for us to follow.

Secondly, I want to thank Prof. Ming Wu and Prof. Junqiao Wu for sitting on my dissertation committee, and Prof. Borivoje Nikolic for serving as a member of my qualifying exam committee. Their feedback and guidance during the qualifying exam was invaluable. And I also appreciate Prof. Muhammad M. Hussein's kindness for giving me the opportunity to TA the graduate solid-state devices course (EE230B).

All the research topics in this dissertation are a direct result of collaborations with many research partners. The invaluable discussions with Yi-Ting Wu are instrumental for me in understanding modern MOSFET concepts. These concepts are fundamental to all of my research topics presented in this dissertation. I am very grateful for his patience and clarity in explaining difficult technical details. The doped iFinFET SRAM project would not have come to fruition without Dr. Daniel Connelly and Dr. Peng Zheng. Their expertise in electronics has helped me clear many technical hurdles. I appreciate Prof. Hiu-Yung Wong's guidance and support on the self-heating project. I would also like to thank Prof. Nattapol Damrongplasit and Dr. Nuo Xu for their research guidance in the last two yearsof my undergraduate study. Members from Atomera Inc., especially Hideki Takeuchi, Dr. Robert Mears, also have offered constructive feedback in my research. In addition, I appreciate the help from the staff members (Joanna Bettinger, Dr. Allison Dove, Richelieu Hemphill, Ryan Rivers, and Dr. William Flounders) from Marvell Nanolab at UC Berkeley during my short stay there working on transistor fabrication.

I also want to thank all other current and graduated members in Prof. Liu's group, in particular, Dr. Sergio Almeida, Hoonsung Choi, Tsegereda Esatu, Xiaoer Hu, Prof. Sangwan Kim, Chanmin Lee, Benjamin Osoba, Dr. Chuang Qian, Dr. Thomas Rembert, Dr. Kimihiko Sato, Urmita Sikder, Lars Tatum, Alice Ye, and Dr. Xi Zhang, for many interesting discussions in research and life topics. In addition, I would also like to thank many current and previous "sitting" members of Cory 373: Qiutong Jin, Yafei Li, Dr. Ruonan Liu, Dr. Jalal Nilchi, Dr. Alper Ozgurluk, Kieran Peleaux, Qianyi Xie. They have made Cory 373 a wonderful workplace.

In addition to the above, I have also received much help from the administrative staff in the EECS department. In particular, I would like to thank Shirley Salanio, Charlotte Jones, and Nicole Song for their support in handling administrative matters.

Finally, I would like to sincerely express my gratitude to my family. The Ph. D. experience would not have been complete without confusion, frustration, and self-doubt from time to time. It is their ceaseless support that has emboldened me to live through many dark moments during these years. Thank you!

Chapter 1

Introduction

1.1 IC Chip Area Scaling – The Central Theme

1.1.1 Moore's Law and Dennard's Scaling Law

In 1965, Gordon Moore observed that the number of transistors on the most advanced integrated circuit "chip" roughly doubled every year; this trend eventually became known as Moore's law [1]. In practice, this is almost equivalent to steady doubling of transistor density because the area of a chip cannot be increased much due to manufacturing cost considerations [2]. Therefore, Moore's law implicitly requires the transistor area to be roughly halved with each new generation ("node") of manufacturing process technology. In 1974, Robert Dennard and his colleagues proposed the constant-field scaling rule, which serves as a set of guidelines for scaling metal oxide semiconductor field-effect transistors (MOSFETs) [3]. In this rule, also known as the Dennard's scaling law, various transistor design parameters are scaled according to a factor (α) so that the peak electric field within the transistor is kept relatively constant across technology generations.

	Constant Field Scaling
Physical dimensions: L_{gate}, W, T_{ox}, wire pitch	$1/\alpha$
Body doping concentration	α
Voltage	$1/\alpha$
Circuit density	$1/\alpha^2$
Capacitance per circuit	$1/\alpha$
Circuit speed	α
Circuit power	$1/\alpha^2$
Power density	1
Power-delay product	$1/\alpha^2$

Table 1.1-1. Dennard scaling law for transistors. Adapted from [4].

As shown in Table 1.1-1, following the constant field scaling approach, circuit performance is increased at a rate of α as we scale the transistor area by $1/\alpha^2$.

However, the industry stopped following the Dennard's scaling law around 2005 because of practical limits and non-ideal effects that become significant as the transistor lateral dimensions are scaled down. For example, Dennard's scaling law requires the body doping concentration to be increased by the scaling factor, α. But as the doping concentration increases, impurity-induced carrier scattering increases and degrades carrier mobility and hence transistor on-state current. In addition, the scaling of physical gate length also started to meet fabrication challenges. More importantly, gate length scaling is limited by short-channel effects (SCE) in conventional planar MOSFET structures. Poor SCE also limit reductions in the supply voltage. To mitigate SCE at sub-30nm gate lengths, FinFETs were introduced.

1.1.2 Challenges for Area Scaling in the FinFET Era and Beyond (Post-Dennard Scaling)

The FinFET has been the transistor design of choice in the semiconductor industry since Intel's 22nm technology node [5] and other foundry's 16/14nm technology node [6] thanks to its superior electrostatic integrity as compared to the planar MOSFET. However, FinFETs are much more difficult to fabricate, which results in increasing wafer cost.

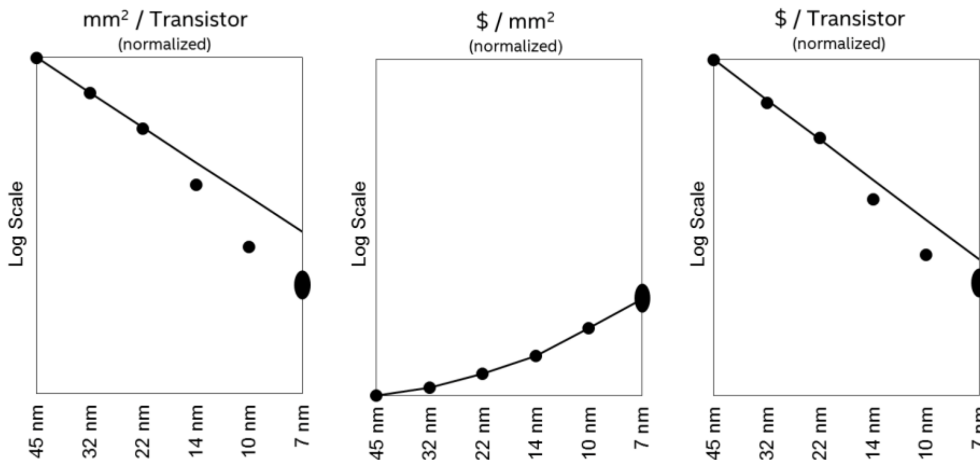


Figure 1.1-1. Fabrication cost increases at a fast pace in small technology nodes. Adapted from [7].

Figure 1.1-1 illustrates the trend of increasing cost per area due to more complicated fabrication process. From a business standpoint, in order to justify the need for a smaller technology node, the transistor area must be aggressively scaled (shown on the left) so that the effective cost per transistor is reduced. However, scaling the transistor area in FinFET-era becomes even more challenging.

Scaling the area in the length direction requires reduction of the contacted (gate electrode) poly-Si pitch (CPP). Generally speaking, the CPP is defined via photo lithography. In order to reduce CPP, more complicated patterning processes are required, as shown in Figure 1.1-2.

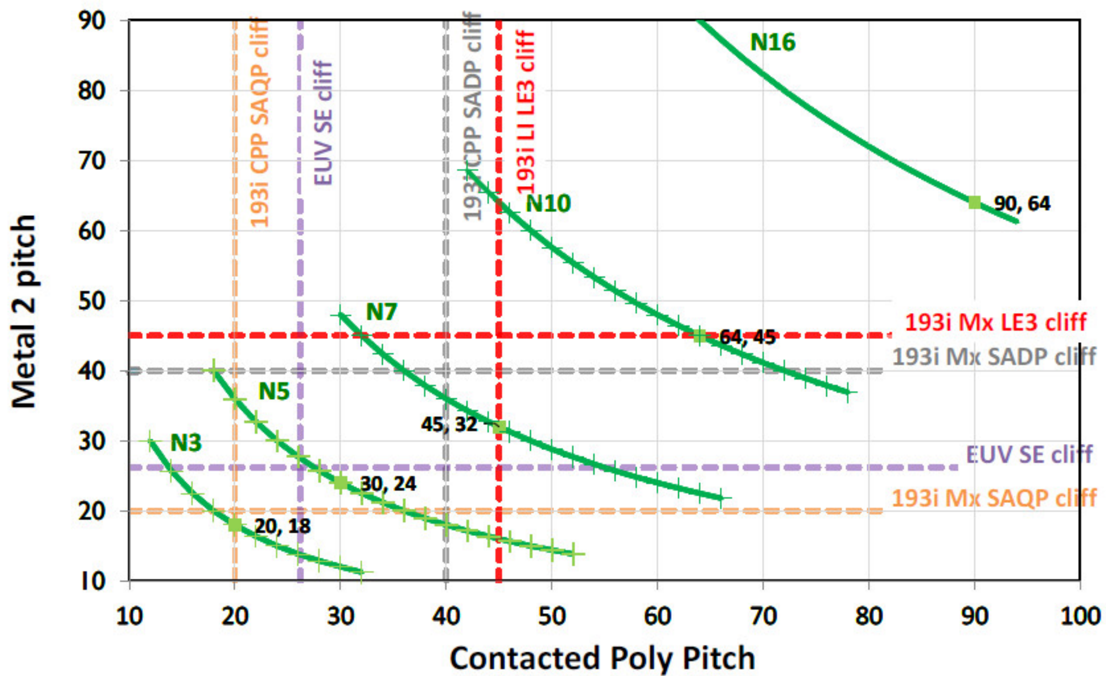


Figure 1.1-2. Limits (cliffs) of different patterning techniques. Adapted from [8]. Note that Metal 2 Pitch (MMP) is directly related to Fin Pitch via gear ratio. Gear ratio is usually 4:3 (=MMP:FP).

More specifically, CPP has 3 components (Figure 1.1-3). L_G is the physical gate length, L_{SP} is the length of sidewall spacer, and L_{SD} is the length of source/drain. As shown in Figure 1.1-4, physical gate scaling has almost stalled in recent generations. On the other hand, reducing gate-sidewall spacer length (L_{SP}) can lead to increase in the gate-to-drain capacitance and may result in transistor reliability issues. In addition, reducing the source/drain length (L_{SD}) can reduce the effectiveness of source/drain stressor, which degrades transistor on-state

current and hence circuit performance. There are also process complications when L_{SD} is reduced.

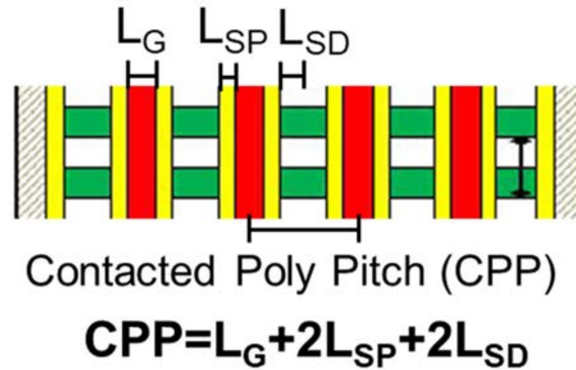


Figure 1.1-3. Definition of CPP. Adapted from [9] with modifications.

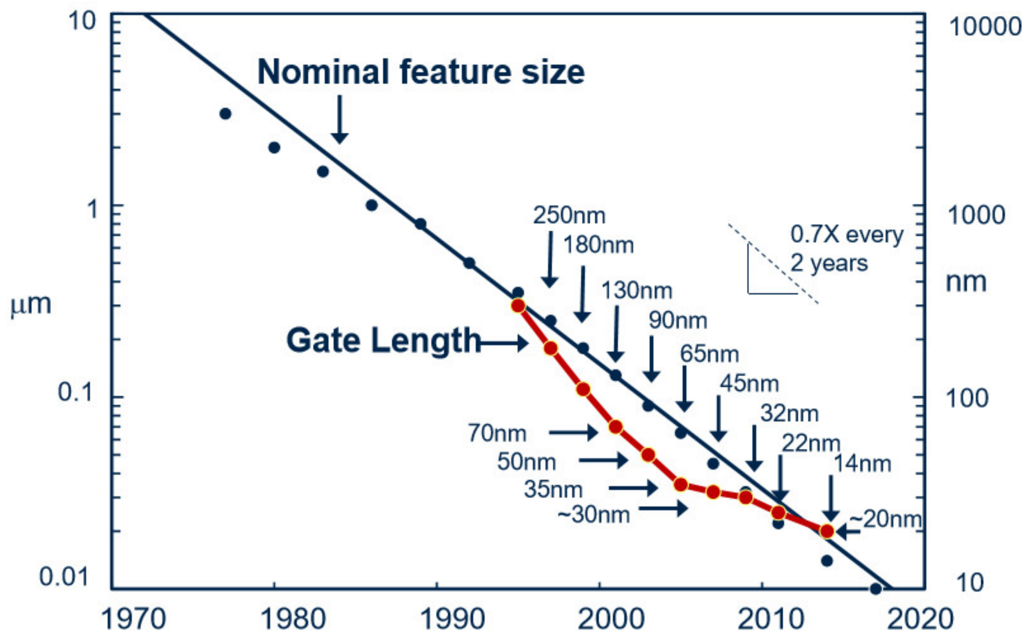


Figure 1.1-4. Physical gate length scaling. Adapted from [10]. Data points are taken from various presentations from Intel Corp..

Scaling in the width direction requires reducing the fin pitch (FP). FP consists of two parts (Figure 1.1-5). W_{FIN} is the fin width and W_{STI} is the width of STI region. Fin width is already sub-10nm and there is not too much room for further reduction. An extremely thin (narrow) fin could lead to performance degradation due to non-idealities like excessive surface carrier-scattering effects. On the other hand, reducing W_{STI} could potentially meet fabrication challenges

due to the need for conformal deposition of oxide to fill trenches with higher aspect ratio.

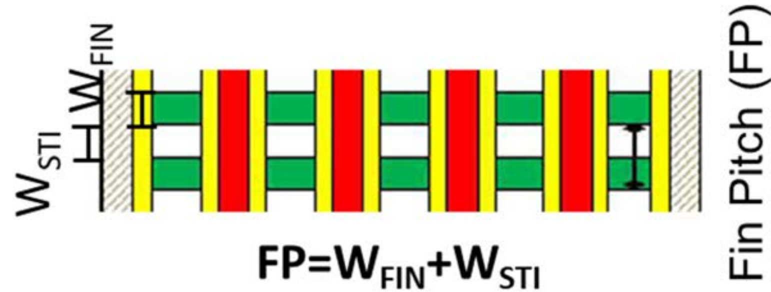


Figure 1.1-5. Definition of FP. Adapted from [9] with modifications.

Table 1.1-2 summarizes the important process parameter values for 3 generations of FinFETs from Intel Corp. [5, 11, 12, 13].

	22nm [5]	**14nm [11]**	**10nm [12, 13]**
CPP (nm)	90	70 (0.78X)	54 (0.77X)
FP (nm)	60	42 (0.7X)	34 (0.81X)
W_{FIN} (nm)	~8	~8	~7
H_{FIN} (nm)	34	42 (1.24X)	46-53 (1.10X-1.25X)

Table 1.1-2. Summary of process parameters in 3 generations of FinFETs from Intel Corp. The number in parentheses is the scaling factor compared to the last generation. For 10nm technology node, a 46nm H_{FIN} was stated in [12] and a 53nm H_{FIN} was found in [13].

1.1.3 Fin-Depopulation for Further Area Scaling

Recognizing the aforementioned challenges of conventional area scaling, the industry has moved to the third dimension (i.e., out of the plane of the silicon wafer surface). In FinFETs, the height of the fin is (almost) directly proportional to the transistor's conductive strength. By increasing the fin height, the current density per layout area can be increased. To meet the same current specification as set by the circuit designers, a smaller number of taller fins is sufficient. As a result, the total layout width can be reduced without significantly adding process complexity. This approach is commonly referred to as "fin depopulation," as

shown in Figure 1.1-6. But the fin height cannot be increased without limitation due to added process complexity and increased parasitic resistance between the metal contacts at the top of the fins in the source/drain regions to the bottom of the fin. To continue to improve the current density, new transistor structures or novel high-mobility channel material are necessary.

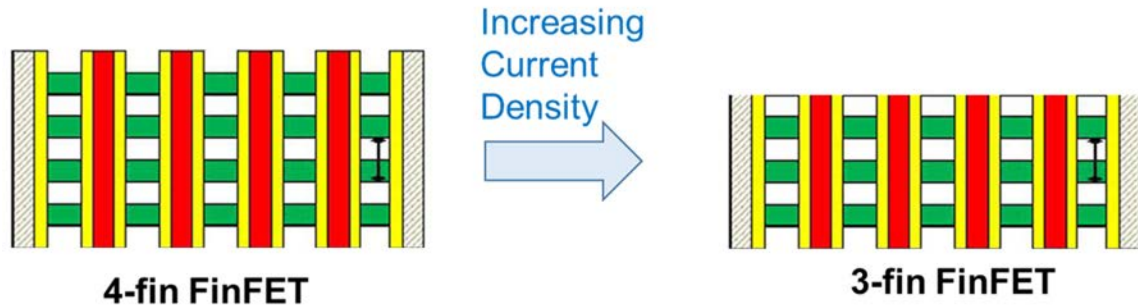


Figure 1.1-6. Fin depopulation. Adapted from [9] with modifications.

1.2 Transistor Design Techniques to Facilitate Further Area Scaling

1.2.1 Advanced Transistor Structures

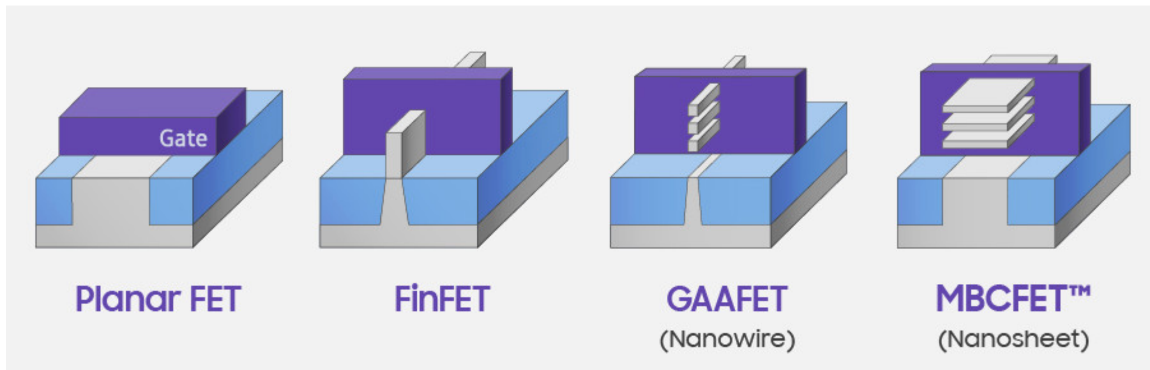


Figure 1.2-1. Advanced transistor structures. Adapted from [14].

Due to excessive short channel effects, conventional bulk Si planar FETs cannot sustain further gate length scaling. Therefore, advanced transistor structures are required. These transistor structures can suppress the short channel effects through better gate control over the channel (Figure 1.2-1).

FinFETs (more specifically, bulk silicon FinFETs) were first introduced in mass production at Intel's 22nm technology node [5]. The rest of the industry then followed. FinFETs are multi-gate transistors in which the gate electrodes cover three sides (left, right, and top) of the fin channel region. This results in better gate control than in planar FETs. In addition, since FinFETs are a variant of thin-body transistors, they feature narrow fins to provide for physical confinement of carriers in the channel region. A smaller fin width is generally preferred for better electrostatic integrity. But FinFET performance might be degraded if the fin is too narrow due to the loss of inversion carrier density and excessive surface roughness scattering [15]. On the other hand, the effective channel width of a FinFET is directly proportional to the outer perimeter of the exposed fin. In high aspect ratio FinFETs, the gate on the top side contributes less so that they are essentially double-gate MOSFET structures.

As mentioned in Section 1.1.3, to facilitate further scaling to sub-5nm technology nodes, novel transistor structures (e.g., Gate-All-Around FET (GAAFET) and Nanosheet FET (NSFET)) are generally expected to be adopted. These advanced structures are expected to achieve even better electrostatic integrity than FinFET.

GAAFETs have gates covering all 4 sides of the nanowire/channel region, which provide for superior electrostatic integrity. Due to the small geometry of the nanowires, in order to meet the drive current requirement, multiple (≥ 3) nanowires must be used. In practice, GAAFETs generally require a large spacing between adjacent nanowires to accommodate the gate stack. Therefore, the aspect ratio of the entire channel stack is much higher than that of the fin in FinFETs, posing potential fabrication challenges. In addition, due to more exposure between the gate and source/drain regions, the parasitic gate-to-drain capacitance [16] is much larger compared to FinFETs.

To address the issues of GAAFETs, NSFETs (also called Multi-Bridge Channel FETs (MBCFETs)) have been proposed. In NSFETs, the gate still covers all 4 sides of the nanosheet/channel. However, since the NSFETs are usually wide, the left and right gates do not contribute much; hence, the electrostatic integrity of NSFETs is between that of FinFETs and GAAFETs. NSFETs can achieve better layout efficiency than FinFETs by eliminating the STI regions between adjacent fins. More importantly, the channel (sheet) width of the NSFETs can be lithographically defined, as opposed to the spacer-defined

width in the case of FinFETs and GAAFETs. This gives more flexibility to circuit designers for adjusting transistor drive strength.

Other evolutionary FinFET structures also have been proposed. For example, the inserted-oxide FinFET (iFinFET) [16] can achieve better electrostatic integrity than the FinFET without adding too much parasitic capacitance. The details of iFinFETs are discussed in **chapter 3**.

1.2.2 High Mobility Channel Materials

The advanced structures described in Section 1.2.1 help with transistor area scaling by mitigating the short channel effects for small gate lengths. In this section, another possibility is discussed: achieving higher current density using high mobility channel materials to enable scaling in the width direction through fin depopulation.

Materials	**Electron Mobility (cm^2/Vs)**	**Hole Mobility (cm^2/Vs)**
Si	1400	470
Ge	3900	1900
GaAs	8500	400
InAs	40000	500
$In_{0.53}Ga_{0.47}As$	12000	300

Table 1.2-1. Summary of electron and hole mobility in conventional high mobility channel materials (bulk).

For n-channel transistors, most III-V alloys such as GaAs, InAs, and InGaAs can achieve very high electron mobility (Table 1.2-1). However, from a process integration perspective, these III-V channels are hardly practical at this point. Due to the large lattice mismatch between the Si and most III-V materials, a thick epitaxial stress relax buffer (SRB) is generally required to gradually reduce the mismatch [17]. Recent advancement in the aspect ratio trapping (ART) technique [18] could facilitate the integration of these exotic materials into a standard CMOS manufacturing process.

The desire for high hole mobility materials is much stronger due to the fact that the embedded SiGe source/drain stressor becomes less effective as we scale

down the volume of these stressors with transistor miniaturization. SiGe alloys and Ge are usually considered the most promising candidates due to their compatibility with a Si-based CMOS manufacturing process. Ge fins can be fabricated on the Si substrate using ART [19], but the Ge fin sidewalls might still suffer from surface passivation problems. On the other hand, low Ge molefraction SiGe becomes attractive as a channel material in p-channel FinFETs due to smaller lattice mismatch with the Si substrate. Even though alloy scattering effects in SiGe cannot be ignored, the combined stress from the Si substrate (or strain-relaxed SiGe buffer layer) and the source/drain regions makes low Ge molefraction SiGe achieve higher hole mobility and hence higher performance than the strained Si [20, 21].

1.3 Research Objectives and Thesis Overview

This dissertation addresses challenges for continued transistor area scaling as follows.

In **chapter 2**, a hetero-channel FinFET design comprising $Si/Si_{1-x}Ge_x/Si$ is evaluated and benchmarked against conventional Si FinFET and $Si_{1-x}Ge_x$ FinFET. In particular, x (i.e., Ge molefraction) is chosen to be small so that $Si_{1-x}Ge_x$ can be directly fabricated on top of a conventional silicon wafer substrate via epitaxial growth.

In **chapter 3**, a cell ratio tuning scheme for iFinFET 6-T SRAM high density cell (HDC) design is proposed. Specifically, the top nanowire(s) can be selectively doped via ion implantation to precisely reduce the transistor drive strength and hence fine tune the bit cell ratio. The feasibility of this approach is validated via a process simulator, Sentaurus Process [22]. The transistor performance and impact of process variations are simulated using a 3-D TCAD software, Sentaurus Device [22]. Finally, the yield of 6-T SRAM is estimated via an in-house developed software [23].

In the first half of **chapter 4**, a comparison of advanced transistor structures with regard to self-heating effects (SHE) is presented. In particular, the performance of FinFETs (FFs), Nanosheet FETs (NSFs), and Nanowire Gate-All-Around FETs (GAAFs) are benchmarked under the constraint of identical peak temperature. To ensure a fair comparison, the design parameters are set for FFs and GAAFs so that their on-state current are similar to that of NSFs. The

difference in SHEs for n-channel *vs.* p-channel transistors is investigated. In addition, the influence of various transistor design parameters on SHE is also investigated. In the end, the operating voltages of NSFs and GAAFs are lowered so that they have the same maximum temperature as that in FFs. Under this scenario, the performance of FFs, NSFs, and GAAFs are compared. In the second half of **chapter 4**, various design parameters are optimized in order to minimize SHE in NSFs.

1.4 References

[1] G. E. Moore, "Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff.," in IEEE Solid-State Circuits Society Newsletter, vol. 11, no. 3, pp. 33-35, Sept. 2006. doi: 10.1109/N-SSC.2006.4785860.

[2] M. Bohr, "A 30 Year Retrospective on Dennard's MOSFET Scaling Paper," in IEEE Solid-State Circuits Society Newsletter, vol. 12, no. 1, pp. 11-13, Winter 2007. doi: 10.1109/N-SSC.2007.4785534.

[3] R. H. Dennard, F. H. Gaensslen, Hwa-Nien Yu, V. L. Rideout, E. Bassous and A. R. Leblanc, "Design Of Ion-implanted MOSFET's with Very Small Physical Dimensions," in Proceedings of the IEEE, vol. 87, no. 4, pp. 668-678, April 1999. doi: 10.1109/JPROC.1999.752522.

[4] W. Haensch, E. J. Nowak, R. H. Dennard, P. M. Solomon, A. Bryant, O. H. Dokumaci, A. Kumar, X. Wang, J. B. Johnson and M. V. Fischetti, "Silicon CMOS devices beyond scaling," in IBM Journal of Research and Development, vol. 50, no. 4.5, pp. 339-361, July 2006. doi: 10.1147/rd.504.0339.

[5] C. Auth, C. Allen, A. Blattner, D. Bergstrom, M. Brazier, M. Bost, M. Buehler, V. Chikarmane, T. Ghani, T. Glassman, R. Grover, W. Han, D. Hanken, M. Hattendorf, P. Hentges, R. Heussner, J. Hicks, D. Ingerly, P. Jain, S. Jaloviar, R. James, D. Jones, J. Jopling, S. Joshi, C. Kenyon, H. Liu, R. McFadden, B. McIntyre, J. Neirynck, C. Parker, L. Pipes, I. Post, S. Pradhan, M. Prince, S. Ramey, T. Reynolds, J. Roesler, J. Sandford, J. Seiple, P. Smith, C. Thomas, D. Towner, T. Troeger, C. Weber, P. Yashar, K. Zawadzki and K. Mistry, "A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM

capacitors," 2012 Symposium on VLSI Technology (VLSIT), Honolulu, HI, 2012, pp. 131-132. doi: 10.1109/VLSIT.2012.6242496S.

[6] S. Y. Wu, C. Y. Lin, M. C. Chiang, J. J. Liaw, J. Y. Cheng, S. H. Yang, S. Z. Chang, M. Liang, T. Miyashita, C. H. Tsai, C. H. Chang, V. S. Chang, Y. K. Wu, J. H. Chen, H. F. Chen, S. Y. Chang, K. H. Pan, R. F. Tsui, C. H. Yao, K. C. Ting, T. Yamamoto, H. T. Huang, T. L. Lee, C. H. Lee, W. Chang, H. M. Lee, C. C. Chen, T. Chang, R. Chen, Y. H. Chiu, M. H. Tsai, S. M. Jang, K. S. Chen and Y. Ku, "An enhanced 16nm CMOS technology featuring 2nd generation FinFET transistors and advanced Cu/low-k interconnect for low power and high performance applications," 2014 IEEE International Electron Devices Meeting, San Francisco, CA, 2014, pp. 3.1.1-3.1.4. doi: 10.1109/IEDM.2014.7046970.

[7] M. T. Bohr, "Logic Technology Scaling to Continue Moore's Law," 2018 IEEE 2nd Electron Devices Technology and Manufacturing Conference (EDTM), Kobe, 2018, pp. 1-3. doi: 10.1109/EDTM.2018.8421433.

[8] J. Ryckaert, P. Raghavan, P. Schuddinck, H. B. Trong, A. Mallik, S. S. Sakhare, B. Chava, Y. Sherazi, P. Leray, A. Mercha, J. Bömmels, G. R. McIntyre, K. G. Ronse, A. Thean, Z. Tökei, A. Steegen and D. Verkest, "DTCO at N7 and beyond: patterning and electrical compromises and opportunities," Proc. SPIE 9427, Design-Process-Technology Co-optimization for Manufacturability IX, 94270C (18 March 2015). doi: 10.1117/12.2178997.

[9] S. K. Marella, A. R. Trivedi, S. Mukhopadhyay and S. S. Sapatnekar, "Optimization of FinFET-based circuits using a dual gate pitch technique," 2015 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), Austin, TX, 2015, pp. 758-763. doi: 10.1109/ICCAD.2015.7372646.

[10] B. Nikolic. (2019). Advanced Digital Integrated Circuits: Lecture 1 – Intro [Powerpoint Slides].

[11] S. Natarajan, M. Agostinelli, S. Akbar, M. Bost, A. Bowonder, V. Chikarmane, S. Chouksey, A. Dasgupta, K. Fischer, Q. Fu, T. Ghani, M. Giles, S. Govindaraju, R. Grover, W. Han, D. Hanken, E. Haralson, M. Haran, M. Heckscher, R. Heussner, P. Jain, R. James, R. Jhaveri, I. Jin, H. Kam, E. Karl, C. Kenyon, M. Liu, Y. Luo, R. Mehandru, S. Morarka, L. Neiberg, P. Packan, A. Paliwal, C. Parker, P. Patel, R. Patel, C. Pelto, L. Pipes, P. Plekhanov, M. Prince, S. Rajamani, J. Sandford, B. Sell, S. Sivakumar, P. Smith, B. Song, K. Tone, T. Troeger, J. Wiedemer, M. Yang and K. Zhang, "A 14nm logic

technology featuring 2nd-generation FinFET, air-gapped interconnects, self-aligned double patterning and a 0.0588 µm2 SRAM cell size," 2014 IEEE International Electron Devices Meeting, San Francisco, CA, 2014, pp. 3.7.1-3.7.3. doi: 10.1109/IEDM.2014.7046976.

[12] C. Auth, A. Aliyarukunju, M. Asoro, D. Bergstrom, V. Bhagwat, J. Birdsall, N. Bisnik, M. Buehler, V. Chikarmane, G. Ding, Q. Fu, H. Gomez, W. Han, D. Hanken, M. Haran, M. Hattendorf, R. Heussner, H. Hiramatsu, B. Ho, S. Jaloviar, I. Jin, S. Joshi, S. Kirby, S. Kosaraju, H. Kothari, G. Leatherman, K. Lee, J. Leib, A. Madhavan and K. Maria, "A 10nm high performance and low-power CMOS technology featuring 3rd generation FinFET transistors, Self-Aligned Quad Patterning, contact over active gate and cobalt local interconnects," 2017 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2017, pp. 29.1.1-29.1.4. doi: 10.1109/IEDM.2017.8268472.

[13] (2017) Intel Technology and Manufacturing Day. [Online]. Available: https://newsroom.intel.com/newsroom/wp-content/uploads/sites/11/2017/03/Kaizad-Mistry-2017-Manufacturing.pdf.

[14] (2019) Samsung Foundry Forum. [Online]. Available: https://news.samsung.com/global/samsung-electronics-leadership-in-advanced-foundry-technology-showcased-with-latest-silicon-innovations-and-ecosystem-platform.

[15] X. He, J. Fronheiser, P. Zhao, Z. Hu, S. Uppal, X. Wu, Y. Hu, R. Sporer, L. Qin, R. Krishnan, E. M. Bazizi, R. Carter, K. Tabakman, A. K. Jha, H. Yu, O. Hu, D. Choi, J. G. Lee, S. B. Samavedam and D. K. Sohn, "Impact of aggressive fin width scaling on FinFET device characteristics," 2017 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2017, pp. 20.2.1-20.2.4. doi: 10.1109/IEDM.2017.8268427.

[16] P. Zheng, D. Connelly, F. Ding and T. J. K. Liu, "FinFET Evolution Toward Stacked-Nanowire FET for CMOS Technology Scaling," in IEEE Transactions on Electron Devices, vol. 62, no. 12, pp. 3945-3950, Dec. 2015. doi: 10.1109/TED.2015.2487367.

[17] R. Oxland, X. Li, S. W. Chang, S. W. Wang, T. Vasen, P. Ramvall, R. Contreras-Guerrero, J. Rojas-Ramirez, M. Holland, G. Doornbos, Y. S. Chang, D. S. Macintyre, S. Thoms, R. Droopad, Y.-C. Yeo, C. H. Diaz, I. G. Thayne and M. Passlack, "InAs FinFETs With $\textrm {H}_{\mathrm {fin}}=20$ nm

Fabricated Using a Top–Down Etch Process," in IEEE Electron Device Letters, vol. 37, no. 3, pp. 261-264, March 2016. doi: 10.1109/LED.2016.2521001.

[18] J. G. Fiorenza, J.-S. Park, J. Hydrick, J. Li, J. Li, M. Curtin, M. Carroll and A. Lochtefel, "Aspect Ratio Trapping: A Unique Technology for Integrating Ge and III-Vs with Silicon CMOS," ECS Trans. 2010 volume 33, issue 6, 963-976. doi: 10.1149/1.3487628.

[19] M.J.H. van Dal, G. Vellianitis, G. Doornbos, B. Duriez, T. M Shen, C. C. Wu, R. Oxland, K. Bhuwalka, M. Holland, T. L. Lee, C. Wann, C. H. Hsieh, B. H. Lee, K. M. Yin, Z. Q. Wu, M. Passlack and C. H. Diaz, "Demonstration of scaled Ge p-channel FinFETs integrated on Si," 2012 International Electron Devices Meeting, San Francisco, CA, 2012, pp. 23.5.1-23.5.4. doi: 10.1109/IEDM.2012.6479089.

[20] C. Jeong, H.-H. Park, S. Dhar, S. Park, K. Lee, S. Jin, W. Choi, U.-H. Kwon, K.-H. Lee and Y. Park, "Physical understanding of alloy scattering in SiGe channel for high-performance strained pFETs," 2013 IEEE International Electron Devices Meeting, Washington, DC, 2013, pp. 12.2.1-12.2.4. doi: 10.1109/IEDM.2013.6724614.

[21] D. Guo, G. Karve, G. Tsutsui, K-Y Lim, R. Robison, T. Hook, R. Vega, D. Liu, S. Bedell, S. Mochizuki, F. Lie, K. Akarvardar, M. Wang, R. Bao, S. Burns, V. Chan, K. Cheng, J. Demarest, J. Fronheiser, P. Hashemi, J. Kelly, J. Li, N. Loubet, P. Montanini, B. Sahu, M. Sankarapandian, S. Sieg, J. Sporre, J. Strane, R. Southwick, N. Tripathi, R. Venigalla, J. Wang, K. Watanabe, C. W. Yeung, D. Gupta, B. Doris, N. Felix, A. Jacob, H. Jagannathan, S. Kanakasabapathy, R. Mo, V. Narayanan, D. Sadana, P. Oldiges, J. Stathis, T. Yamashita, V. Paruchuri, M. Colburn, A. Knorr, R. Divakaruni, H. Bu and M. Khare, "FINFET technology featuring high mobility SiGe channel for 10nm and beyond," 2016 IEEE Symposium on VLSI Technology, Honolulu, HI, 2016, pp. 1-2. doi: 10.1109/VLSIT.2016.7573360.

[22] Sentaurus User’s Manual, Version L-2016.03, Synopsys, Inc., Mountain View, CA, USA.

[23] A. E. Carlson, “Device and circuit techniques for reducing variation in nanoscale SRAM,” Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Univ. California Berkeley, Berkeley, CA, USA, 2008.

Chapter 2

Si/$Si_{1-x}Ge_x$/Si Hetero-Channel FinFET for Enhanced P-Channel Performance in Low-Power Applications

2.1 Introduction

The FinFET structure was adopted for high-volume manufacture of digital integrated circuits beginning at the 22 nm technology node because of (1) its superior electrostatic integrity as compared with the conventional planar MOSFET structure as well as (2) benefits from layout width scaling [1]. As conventional gate length scaling becomes more difficult, alternative approaches to decreasing the layout area of a FinFET, such as fin de-population [2], will be needed.

To increase the drive current per fin and thereby allow for a reduction in the number of fins required, FinFETs incorporating of high-mobility channel materials such germanium (Ge) [3], and III-V compound semiconductors [4] have been investigated. However, these are not as scalable to sub-10 nm gate length (i.e., beyond the 5 nm technology node) as silicon (Si) FinFETs, due to their lighter carrier tunneling effective mass (which results in degraded subthreshold swing) and larger dielectric permittivity (which results in greater drain-induced barrier lowering). Furthermore, these high-mobility channel materials are difficult to grow with low defect density on silicon wafer substrates, due to their lattice mismatch with silicon, even with the aspect ratio trapping technique (ART) [3].

Recently, silicon-germanium (SiGe) FinFET has been demonstrated to have better on-state performance than Si based FinFET [5]. Despite additional alloy scattering [6], a low Ge mole fraction SiGe fin grown on a Si substrate can still provide for higher hole mobility than a Si fin due to enhanced stress [7]. To further improve the performance of a SiGe FinFET, we propose herein a heterogeneous FinFET design in which only the inner portion of the gated fin channel region is replaced by $Si_{1-x}Ge_x$, as illustrated in Figure 2.1-1. Such a hetero-channel structure can be fabricated using a conventional fabrication process flow by starting with a silicon wafer substrate with an epitaxial layer of

$Si_{1-x}Ge_x$, if the Ge mole fraction x is low such that the epitaxial layer is thinner than the critical thickness for strain relaxation to occur. The structure also can be fabricated using aspect ratio trapping (ART) as described in [3].

Figure 2.1-1. Double cutaway views of the two transistor structures studied in this work. (Left: Control Si FinFET, middle: Hetero-Channel FinFET, right: Control SiGe FinFET). The dashed lines circumscribe the hetero-channel. Different colors are used for the $Si_{1-x}Ge_x$ in the channel region vs. the $Si_{1-y}Ge_y$ in the embedded source and drain regions, to denote different Ge mole fractions. In this work, $Si_{0.5}Ge_{0.5}$ is used for the source/drain contact regions. The fin height is defined to be the height of the fin above the shallow trench isolation (STI). Leff, the effective channel length, is defined to be the lateral distance between the locations where the S/D dopant concentration falls to 2×10^{19} cm^{-3}.

2.2 Hetero-Channel ($Si/Si_{0.9}Ge_{0.1}/Si$) FinFET

The hetero-channel FinFET design parameter values used in this work (summarized in Table 2.2-1) are based on the ITRS 6/5nm technology node, for an off-state leakage specification I_{off} = 100 pA/(µm fin pitch) [8]. Due to even tighter feature pitch than described in [9], we expect the epitaxially grown $Si_{0.5}Ge_{0.5}$ source and drain regions to merge between fins, so that they are more box shaped than diamond shaped. The fin width of 7 nm was selected based on the average fin width in [10].

The Ge mole fraction in the channel region is first chosen to be 10% so that the fin height is thinner than the critical thickness [11]. The hetero-channel FinFET can be fabricated using a process similar to that for a conventional bulk-Si FinFET, with the following extra steps: (1) after the source/drain regions are etched back, an isotropic etch is used to laterally recess the $Si_{0.9}Ge_{0.1}$ underneath the gate-sidewall spacers, (2) Si is epitaxially grown to fill in the laterally recessed regions, prior to epitaxial growth of the $Si_{0.5}Ge_{0.5}$ source/drain regions. It should be noted that such an isotropic etch has been proposed for a bulk-Si FinFET process, to increase channel stress [12]. The recess length therefore corresponds to the Si thickness in the hetero-channel region. See appendix for the more detailed description on process flow.

To achieve an equivalent oxide thickness (EOT) of 0.6 nm, the gate dielectric comprises a 0.4 nm-thick interfacial layer (i.e., SiO_x for the control Si FinFET and $Si_{0.9}Ge_{0.1}O_x$ for the hetero-channel FinFET) and a 1.28 nm-thick high-permittivity layer (k = 25). The nominal supply voltage (V_{DD}) is 0.65 V and the drain voltage for linear operation (V_{DLIN}) is 50 mV. Sentaurus Process [13] is used to model stress inside the transistor structures.

Transistor performance is simulated using Sentaurus Device with the drift-diffusion transport model, ballistic mobility model, density gradient quantum correction model with orientation dependent coefficients, stress-dependent mobility model, and band-to-band tunneling model [13]. Such models have been calibrated to non-equilibrium Green function (NEGF) simulations [14]. The gate-sidewall spacer (k = 5) length, source/drain region length, and punch-through stopper location are co-optimized to provide for the highest on-state current in the control Si FinFET for the given I_{off} specification, within the constraints of gate pitch and fin pitch. The same design parameter values are used for the hetero-channel FinFET and $Si_{0.9}Ge_{0.1}$ FinFET.

	Si FinFET	**Hetero-Channel FinFET**	**$Si_{0.9}Ge_{0.1}$ FinFET**
Gate Pitch (nm)	32		
Fin Pitch (nm)	20		
Gate Length (nm)	12		
Low-k Spacer Length (nm)	5		
Raised S/D Height (nm)	3		
Recess Length (nm)	N/A	1/2/3/4	N/A
EOT (nm)	0.6		
Fin Width (nm)	7		
Fin Height (nm)	46		

Table 2.2-1. Design parameter values for transistors. Note that as a corner case, a hetero-channel FinFET with 0nm recess length is essentially the same as the SiGe FinFET since there are no regrown Si regions near the source and drain regions. EOT is the equivalent oxide thickness, which is calculated by adding the interfacial layer thickness (i.e., 0.4nm) to the equivalent SiO_2 thickness of the high-k HfO_2 layer (i.e., 0.2nm = 1.28nm/25×3.9). Due to the low Ge molefraction (i.e., 0.1) used in the designs, we assume the $Si_{0.9}Ge_{0.1}O_x$ has the same permittivity as SiO_2.

Figure 2.2-1 shows the longitudinal stress profiles and average values within each semiconductor region. Hetero-channel FinFETs has higher, in magnitude, average longitudinal compressive stress underneath the gate (i.e., $Si_{0.9}Ge_{0.1}$ region) due to additional stress from the underlying Si substrate as well as the regrown Si region. The $Si_{0.9}Ge_{0.1}$ FinFET also experiences higher (in magnitude) longitudinal compressive stress in the channel also due to the fact that the lattice mismatch comes from the source/drain $Si_{0.5}Ge_{0.5}$ regions and the bottom Si substrate.

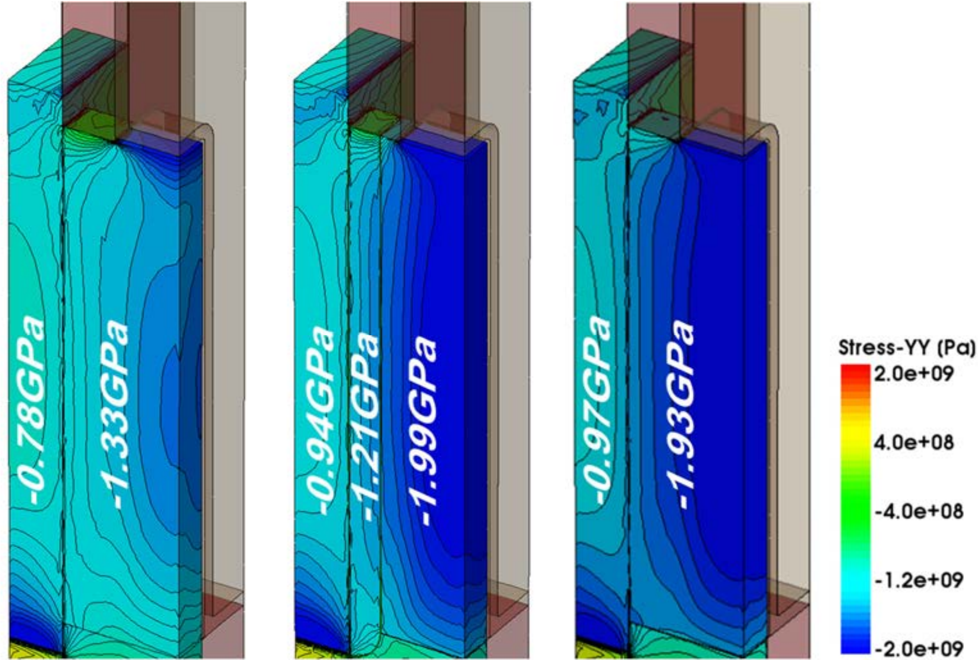


Figure 2.2-1. Longitudinal stress contours from S-Process. (Left: Si FinFET, middle: Hetero-Channel FinFET, right: $Si_{0.9}Ge_{0.1}$ FinFET). A stress dependent mobility model is used in the subsequent S-Device 3-D TCAD simulations. The average longitudinal stress in S/D and channel are annotated over the corresponding regions.

Figure 2.2-2 shows the simulated FinFET transfer and output characteristics, and Table 2.2-2 provides a summary comparison of key performance parameters. The enhanced performance of the hetero-channel FinFET stems from the valence band (VB) offset between $Si_{0.9}Ge_{0.1}$ and Si along two directions: In the vertical (depth) direction, the VB offset provides for lower sub-threshold leakage because it poses a barrier for holes to enter the Si sub-fin region. Figure 2.2-3 shows the energy band diagram along the channel direction: The regrown Si regions underneath the gate-sidewall spacers, corresponding to the shaded regions in Figure 2.2-3, have higher hole potential energy, as compared to $Si_{0.9}Ge_{0.1}$. Hence, as shown in Figure 2.2-3 (b), in a (p-channel) hetero-channel FinFET, holes see increased source-side diffusion barrier in the off-state when compared with the control $Si_{0.9}Ge_{0.1}$ FinFET (i.e., a baseline "hetero-channel" FinFET with zero recess length case). The source side barrier in the hetero-channel FinFET is the same as that in the case of Si FinFET. This effect is similar to that of the "halo" implant in a planar MOSFET [15].

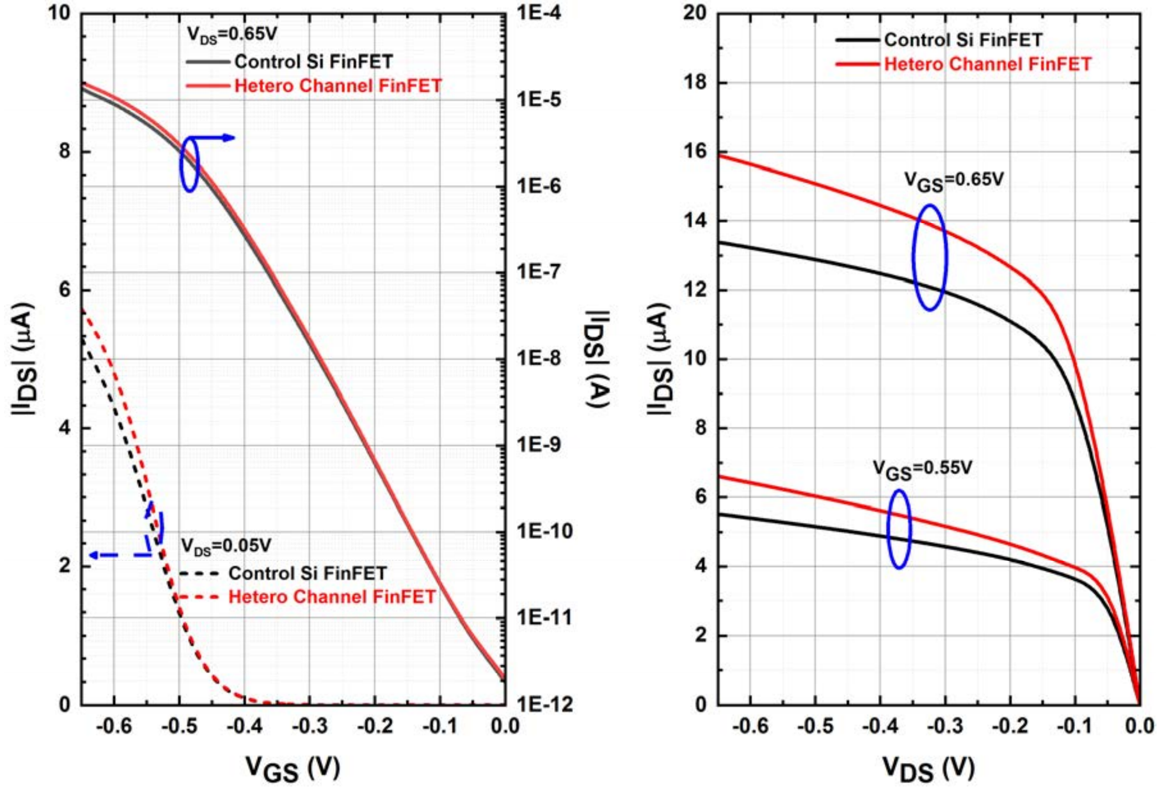
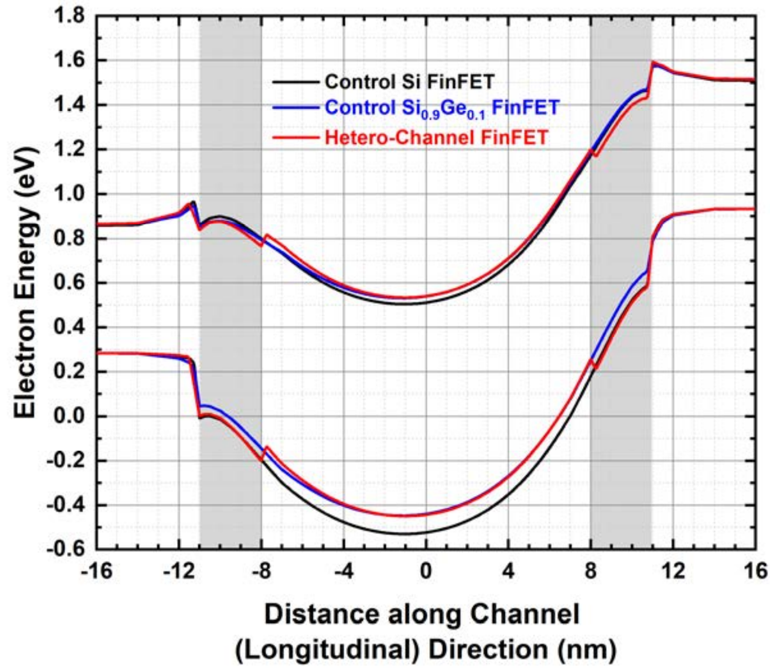


Figure 2.2-2. Simulated transfer (left) and output (right) characteristics for control Si FinFET and hetero-channel FinFET. Output characteristics are simulated at V_{GS} = 0.55V and 0.65V.

To decouple the effects of the two valance band offsets, results for the aforementioned control $Si_{0.9}Ge_{0.1}$ FinFET are also included in the third column of Table 2.2-2. A comparison shows that the vertical VB offset near the top of the STI accounts for 9% of the improvement in on-state current, while the longitudinal VB offset accounts for an additional 7%.



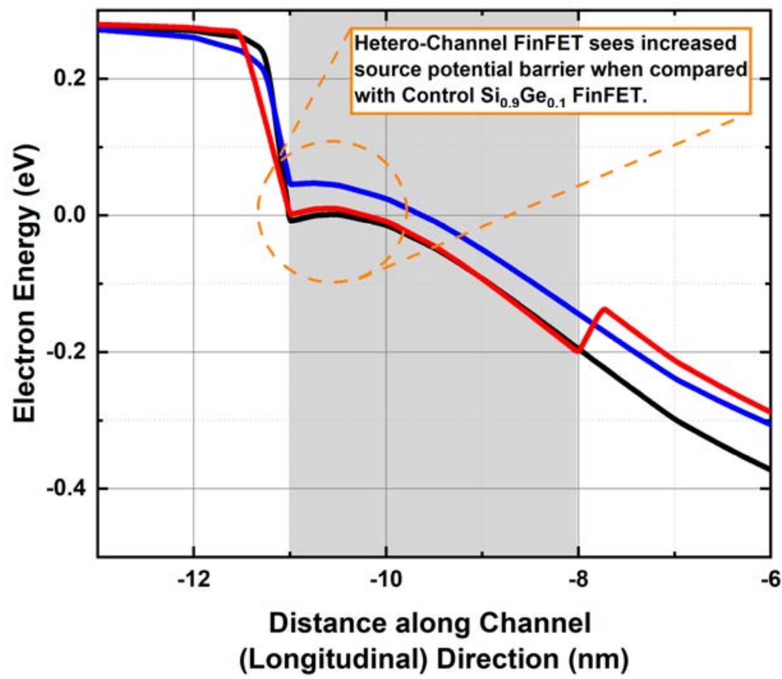


Figure 2.2-3. Band diagram along channel (longitudinal) direction. The shaded region indicates Si (higher bandgap) material in the hetero-channel FinFET. Band diagrams of control Si FinFET and SiGe FinFET (with same molefraction as in channel SiGe in hetero-channel) are also included for reference.

Metric	**Si FinFET**	**Hetero-Channel FinFET (Recess Length = 2nm)**	**$Si_{0.9}Ge_{0.1}$ FinFET**
$\|I_{off}\|$ (pA)	2.0		
$\|I_{on}\|$ (uA)	13.8	16.0 (+16%)	15.1 (+9%)
$\|I_{eff}\|$ (uA)	5.98	6.91 (+16%)	6.62 (+11%)
$V_{t,sat}$ (V)	-0.36	-0.36	-0.37
$V_{t,lin}$ (V)	-0.40	-0.40	-0.41
DIBL (mV/V)	67	67	67
SS_{sat} (mV/dec)	74	73	75
C_{gg} (aF)	44.6	44.2	43.7
$C_{gg}\|V_{gg}\|/\|I_{on}\|$ (ps)	2.11	1.80 (-15%)	1.89 (-11%)

Table 2.2-2. Comparisons of transistor performance metric (V_{dsat}=V_{dd}=0.65V, V_{dlin}=0.05V). $^{a}V_{T}$ is extracted at 100nA×W_{pitch}/L_{eff} = 0.113uA. W_{pitch} is the fin pitch. The percentage in parentheses is relative to the control Si FinFET result.

In general, hetero-channel FinFETs with larger recess length exhibit steeper subthreshold swing. However, there exists a design tradeoff between improved electrostatic integrity and increased on-state resistance, for the recess length. Figure 2.2-4 shows how the on/off current ratio varies with recess length. The optimal design, with recess length = 3 nm, has similar subthreshold swing and drain-induced barrier lowering values as for the control Si FinFET.

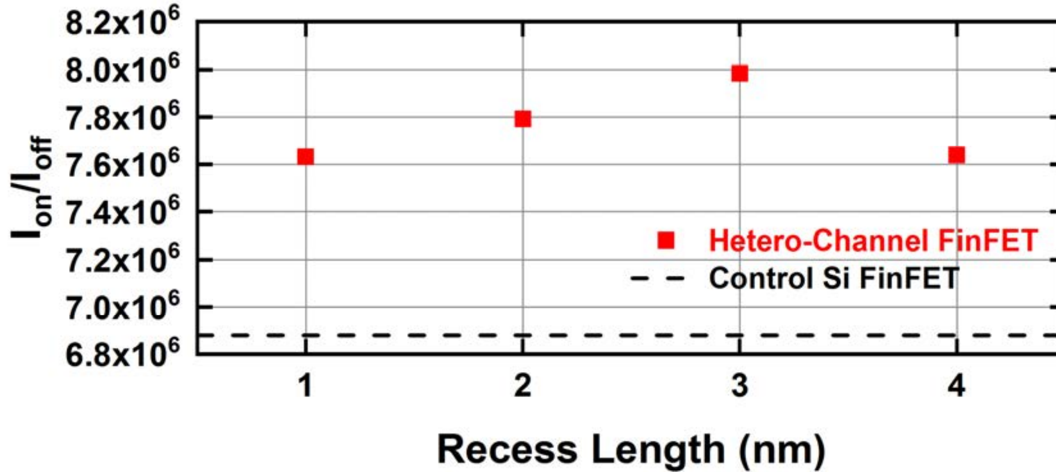


Figure 2.2-4. Impact of recess length (in nm) on the performance of hetero-channel FinFET.

There is little difference in total gate capacitance (C_{gg}) since a relatively low Ge mole fraction is used in the channel region. Therefore, the intrinsic delay is reduced by 15% for the optimized hetero-channel FinFET vs. the control Si FinFET.

2.3 Hetero-Channel FinFETs with Varying Ge Molefraction

In Section 2.2, the hetero-channel design featuring $Si_{0.9}Ge_{0.1}$ is presented. In this section, we extend this study to different Ge molefractions. Modern FinFETs generally require a fin height of more than 40nm. As shown in Figure 2.3-1, in order to grow $Si_{1-x}Ge_x$ thick enough for fin height directly on top of the silicon substrate, the Ge molefraction should be less than 0.25 so that the critical thickness of $Si_{1-x}Ge_x$ is above 100nm. Otherwise, complicated processes such as ART might be required.

In this study, the Ge molefraction is varied from 0.05 to 0.25, while the high bandgap channel region is fixed to silicon. Table 2.3-1 summarizes the performance of different hetero-channel FinFET designs featuring different Ge molefraction $Si_{1-x}Ge_x$/Si channels.

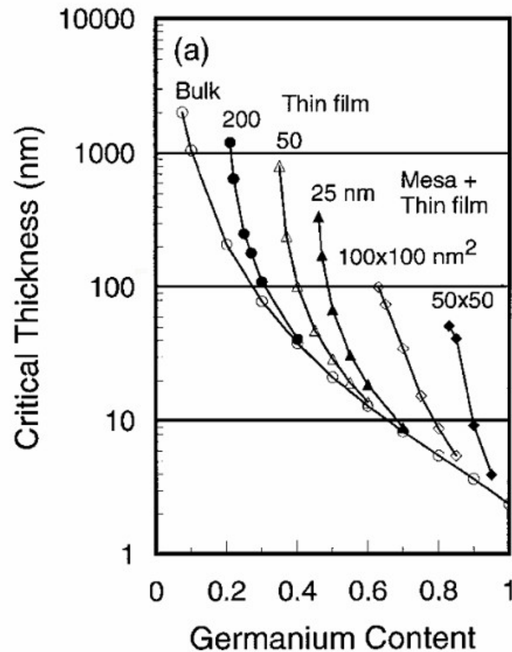


Figure 2.3-1. Critical thickness of $Si_{1-x}Ge_x$ alloy as a function of germanium content on different substrates. Adapted from [11].

From Table 2.3-1, it is shown that with increasing Ge molefraction, the control SiGe FinFET has worse electrostatic integrity due to smaller bandgap and higher permittivity. However, the on-state current of control SiGe FinFET can be steadily improved due to increased compressive stress in the channel. For hetero-channel FinFET designs, the best performance case always happens when the Si channel region is 3nm thick for all Ge molefractions under study. With the Ge molefraction being 0.15 or 0.2, the best hetero-channel design can achieve +17% improvement over the control Si FinFET. However, as the Ge molefraction increases, the sensitivity of performance to the recess length becomes much larger. For example, when the Ge molefraction is 0.25, if the recess length is 4nm, the on-state current of the hetero-channel FinFET is only 55% of that of the control Si FinFET. This is due to the large bandgap difference between Si and $Si_{0.75}Ge_{0.25}$ region and in the on-state, hole injection becomes more difficult and hence degrades the transistor performance. Therefore, considering the performance sensitivity, the optimal Ge molefraction is 0.1.

xGeFin	Recess Length (nm)	|Ion| (µA)	%Ion	SS_{SAT} (mV/dec)
0	Si-FF	13.8	100%	74
0.05	4	15.2	110%	74
	3	15.4	112%	73
	2	15.1	109%	74
	1	14.9	108%	74
	SiGe-FF	14.5	105%	74
0.1	4	15.3	111%	73
	3	15.9	116%	73
	2	15.6	113%	74
	1	15.2	110%	75
	SiGe-FF	15.0	109%	75
0.15	4	14.1	102%	73
	3	16.2	117%	73
	2	15.7	114%	74
	1	15.4	112%	75
	SiGe-FF	15.3	111%	75
0.2	4	11.0	80%	73
	3	16.2	117%	73
	2	15.8	114%	74
	1	15.3	111%	76
	SiGe-FF	15.4	112%	76
0.25	4	7.6	55%	73
	3	15.8	114%	73
	2	15.5	112%	75
	1	15.1	109%	76
	SiGe-FF	15.4	112%	76

Table 2.3-1. Performance summary of hetero-channel designs featuring varying Ge molefraction SiGe. xGeFin is the Ge molefraction used for $Si_{1-x}Ge_x$ channel in both the control SiGe FinFET and the hetero-channel FinFETs. "Si-FF" represents the control Si channel FinFET. ""SiGe-FF" in the second column represents the control SiGe channel FinFET with the same Ge molefraction as in the corresponding hetero-channel designs. SiGe-FF can be thought of hetero-channel design with 0nm recess length.

2.4 Conclusion

In Section 2.2, the performance of a p-channel FinFET comprising a heterogeneous silicon (Si) and silicon-germanium ($Si_{0.9}Ge_{0.1}$) channel region is evaluated using three-dimensional (3-D) TCAD simulations, and benchmarked against a conventional p-channel Si FinFET and SiGe FinFET (with the same Ge content). The results show that the optimal hetero-channel design provides for larger on-state current while maintaining comparable electrostatic integrity as the conventional design due to the valence band (VB) offset between $Si_{0.9}Ge_{0.1}$ and Si. The enhanced performance is achieved with relatively low Ge mole fraction (10%) in the channel region, for ease of manufacture.

In Section 2.3, hetero-channel designs featuring varying Ge molefraction (x = 0 - 0.25) are investigated. It is shown that at x = 0.15 or 0.2, even though the best cases (i.e., recess length = 3nm) can achieve better performance than that in x = 0.1, the performance sensitivity to the recess length becomes much larger, making them less favorable for fabrication.

Therefore, the hetero-channel FinFET featuring $Si/Si_{0.9}Ge_{0.1}/Si$ is a promising candidate for future low-power applications.

2.5 Appendix: Process Simulation in Sentaurus Process

Since the hetero-channel consists of two different materials (i.e., Si and SiGe), it is important to model the channel stress correctly. And the final stress in the channel depends on the ordering of processing steps in the process flow. Therefore, the process simulator, Sentaurus Process (S-Process), is used to construct the hetero-channel FinFET structure. The Sentaurus advanced calibration is also turned on to accurately model the stress [16].

Figure 2.5-1 shows the process flow for constructing the hetero-channel FinFETs. This process is largely similar to the conventional bulk Si FinFET, except for step 1 and 7. In step 1, since only a low Ge molefraction SiGe is used, it can be deposited directly on top of the silicon substrate via epitaxy growth. Aspect ratio trapping (ART) growth technique is not required. In step 7, after the fin recess to make room for the SiGe source/drain regions, an isotropic etching process is performed, followed by the Si epitaxy. This Si refills the vacancy left by the prior isotropic etching process and form the high bandgap region as required in the hetero-channel design. Afterwards, the embedded

source/drain is grown via SiGe epitaxy. The rest of the process (e.g., replacement metal gate) stays the same as in a Si FinFET process.

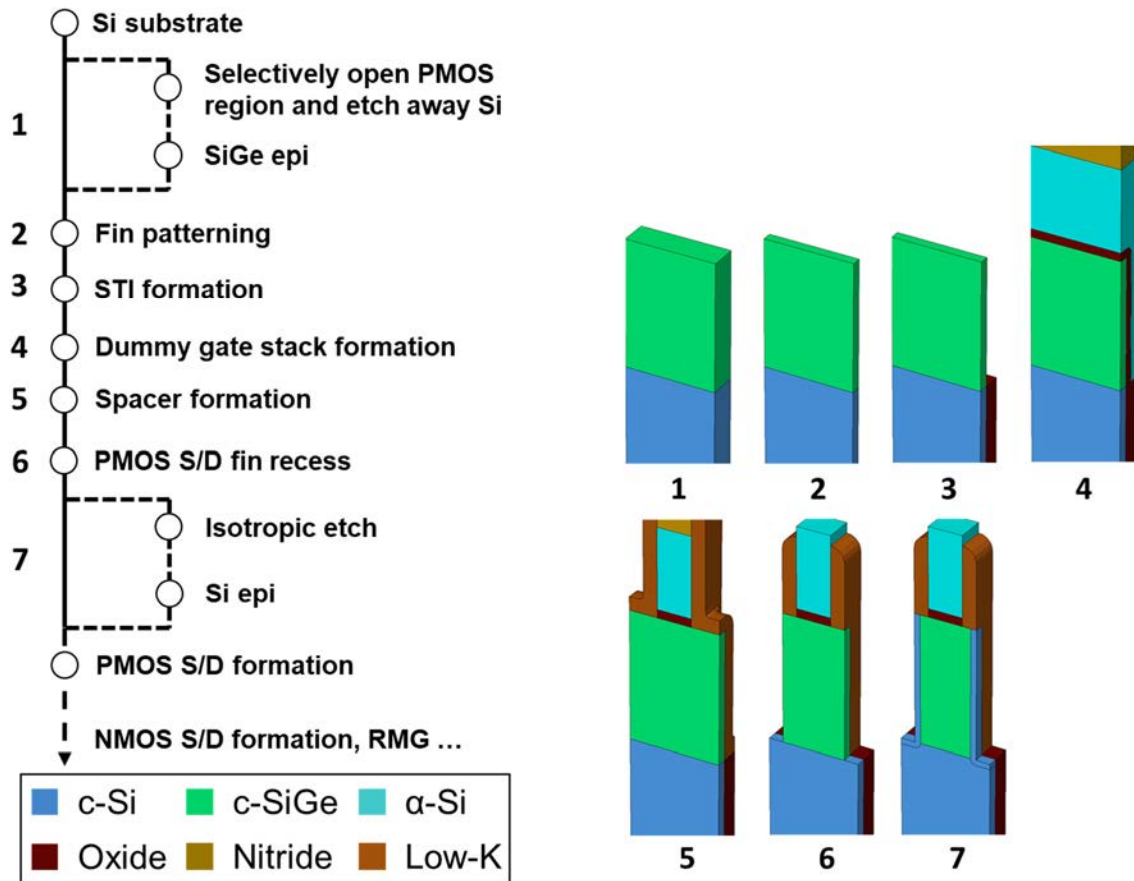


Figure 2.5-1. The process flow in S-Process to construct the hetero-channel FinFETs. S/D epitaxial growth for p-channel MOSFETs and subsequent replacement metal gate (RMG) process are not shown since they are the same as in the conventional bulk Si FinFET process.

2.6 References

[1] C. Auth, C. Allen, A. Blattner, D. Bergstrom, M. Brazier, M. Bost, M. Buehler, V. Chikarmane, T. Ghani, T. Glassman, R. Grover, W. Han, D. Hanken, M. Hattendorf, P. Hentges, R. Heussner, J. Hicks, D. Ingerly, P. Jain, S. Jaloviar, R. James, D. Jones, J. Jopling, S. Joshi, C. Kenyon, H. Liu, R. McFadden, B. McIntyre, J. Neirynck, C. Parker, L. Pipes, I. Post, S. Pradhan, M. Prince, S. Ramey, T. Reynolds, J. Roesler, J. Sandford, J. Seiple, P. Smith, C. Thomas, D. Towner, T. Troeger, C. Weber, P. Yashar, K. Zawadzki and K.

Mistry, "A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors," 2012 Symposium on VLSI Technology (VLSIT), Honolulu, HI, 2012, pp. 131-132. doi: 10.1109/VLSIT.2012.6242496S.

[2] J. Ryckaert, P. Raghavan, P. Schuddinck, H. B. Trong, A. Mallik, S. S. Sakhare, B. Chava, Y. Sherazi, P. Leray, A. Mercha, J. Bömmels, G. R. McIntyre, K. G. Ronse, A. Thean, Z. Tökei, A. Steegen and D. Verkest, "DTCO at N7 and beyond: patterning and electrical compromises and opportunities," Proc. SPIE 9427, Design-Process-Technology Co-optimization for Manufacturability IX, 94270C (18 March 2015). doi: 10.1117/12.2178997.

[3] M. J. H. van Dal, G. Vellianitis, B. Duriez, G. Doornbos, C.-H. Hsieh, B.-H. Lee, K.-M. Yin, M. Passlack and C. H. Diaz, "Germanium p-Channel FinFET Fabricated by Aspect Ratio Trapping," in IEEE Transactions on Electron Devices, vol. 61, no. 2, pp. 430-436, Feb. 2014. doi: 10.1109/TED.2013.2295883.

[4] A. V. Thathachary, G. Lavallee, M. Cantoro, K. K. Bhuwalka, Y.-C. Heo, S. Maeda and S. Datta, "Impact of Sidewall Passivation and Channel Composition on InxGa1-xAs FinFET Performance," in IEEE Electron Device Letters, vol. 36, no. 2, pp. 117-119, Feb. 2015. doi: 10.1109/LED.2014.2384280.

[5] D. Guo, G. Karve, G. Tsutsui, K-Y Lim, R. Robison, T. Hook, R. Vega, D. Liu, S. Bedell, S. Mochizuki, F. Lie, K.Akarvardar, M. Wang, R. Bao, S. Burns, V. Chan, K. Cheng, J. Demarest, J. Fronheiser, P. Hashemi, J. Kelly, J. Li, N. Loubet, P. Montanini, B. Sahu, M. Sankarapandian, S. Sieg, J. Sporre, J. Strane, R. Southwick, N. Tripathi, R. Venigalla, J. Wang, K. Watanabe, C. W. Yeung, D. Gupta, B. Doris, N. Felix, A. Jacob, H. Jagannathan, S. Kanakasabapathy, R. Mo, V. Narayanan, D. Sadana, P. Oldiges, J. Stathis, T. Yamashita, V. Paruchuri, M. Colburn, A. Knorr, R. Divakaruni, H. Bu and M. Khare, "FINFET technology featuring high mobility SiGe channel for 10nm and beyond," 2016 IEEE Symposium on VLSI Technology, Honolulu, HI, 2016, pp. 1-2. doi: 10.1109/VLSIT.2016.7573360.

[6] M. V. Fischetti and S. E. Laux, "Band structure, deformation potentials, and carrier mobility in strained Si, Ge, and SiGe alloys," Journal of Applied Physics, 1996, 80:4, 2234-2252, doi: 10.1063/1.363052.

[7] C. Jeong, H.-H. Park, S. Dhar, S. Park, K. Lee, S. Jin, W. Choi, U.-H. Kwon, K.-H. Lee and Y. Park, "Physical understanding of alloy scattering in SiGe channel for high-performance strained pFETs," 2013 IEEE International Electron

Devices Meeting, Washington, DC, 2013, pp. 12.2.1-12.2.4. doi: 10.1109/IEDM.2013.6724614.

[8] (2015) International Technology Roadmap for Semiconductors (ITRS). [Online]. Available: http://www.itrs2.net/itrs-reports.html.

[9] G. Enemana, D. P. Bruncoa, L. Wittersa, B. Vincenta, P. Faviaa, A. Hikavyya, A. D. Keersgietera, J. Mitarda, R. Looa, A. Velosob, O. Richarda, H. Bendera, W. Vandervorsta, M. Caymaxa, N. Horiguchib, N. Collaerta and A. Theana, "Stress Simulations of Si- and Ge-Channel FinFETs for the 14 nm-Node and Beyond," ECS Trans. 2013 volume 53, issue 1, 225-236. doi: 10.1149/05301.0225ecst.

[10] C. Auth, A. Aliyarukunju, M. Asoro, D. Bergstrom, V. Bhagwat, J. Birdsall, N. Bisnik, M. Buehler, V. Chikarmane, G. Ding, Q. Fu, H. Gomez, W. Han, D. Hanken, M. Haran, M. Hattendorf, R. Heussner, H. Hiramatsu, B. Ho, S. Jaloviar, I. Jin, S. Joshi, S. Kirby, S. Kosaraju, H. Kothari, G. Leatherman, K. Lee, J. Leib, A. Madhavan and K. Maria, "A 10nm high performance and low-power CMOS technology featuring 3rd generation FinFET transistors, Self-Aligned Quad Patterning, contact over active gate and cobalt local interconnects," 2017 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2017, pp. 29.1.1-29.1.4. doi: 10.1109/IEDM.2017.8268472.

[11] F. Y. Huang, "Theory of Strain Relaxation for Epitaxial Layers Grown on Substrate of a Finite Dimension,", in Phys. Rev. Lett., vol. 85, no. 4, pp. 784-787, July. 2000. doi: 10.1103/PhysRevLett.85.784.

[12] M. Garcia Bardon, V. Moroz, G. Eneman, P. Schuddinck, M. Dehan, D. Yakimets, D. Jang, G. Van der Plas, A. Mercha, A. Thean, D. Verkest and A. Steegen, "Layout-induced stress effects in 14nm & 10nm FinFETs and their impact on performance," 2013 Symposium on VLSI Circuits, Kyoto, 2013, pp. T114-T115.

[13] Sentaurus User's Manual, Version L-2016.03, Synopsys, Inc., Mountain View, CA, USA.

[14] M. Choi, V. Moroz, L. Smith and J. Huang, "Extending drift-diffusion paradigm into the era of FinFETs and nanowires," 2015 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), Washington, DC, 2015, pp. 242-245. doi: 10.1109/SISPAD.2015.7292304.

[15] C. Gupta, H. Agarwal, S. Dey, C. Hu and Y. S. Chauhan, "Analysis and modeling of capacitances in halo-implanted MOSFETs," 2017 IEEE Electron Devices Technology and Manufacturing Conference (EDTM), Toyama, 2017, pp. 198-200. doi: 10.1109/EDTM.2017.7947563.

[16] Advanced Calibration for Process Simulation User Guide, Version L-2016.03, Synopsys, Inc., Mountain View, CA, USA.

Chapter 3

Cell Ratio Tuning for High-Density SRAM Voltage Scaling with Inserted-Oxide FinFETs (iFinFETs)

3.1 Introduction

Today bulk-silicon FinFET technology is used for high-volume manufacturing of sub-20 nm-generation CMOS integrated circuits [1, 2]. Due to the {110} fin-sidewall (channel surface) crystallographic orientation and a higher level of mechanical strain induced in the channel region by embedded silicon-germanium source/drain regions, the drive strength of a p-channel (PMOS) FinFET is comparable to that of an n-channel (NMOS) FinFET of the same fin height [1]. Although this may be favorable for high-speed digital logic applications, it results in poor write-ability of a minimally sized six-transistor (6-T) static memory (SRAM) cell comprising two single-fin PMOS pull-up (PU) FinFETs, two single-fin NMOS pass-gate (PG) FinFETs, and two single-fin NMOS pull-down (PD) FinFETs [3]. The drive strength of the PU FinFETs can be reduced by making their effective fin height shorter; however, methods such as selectively adjusting the physical height of the gated fin by selectively adjusting the recess depth of the shallow trench isolation (STI) oxide [4] or by adjusting the depth of the punchthrough-stopper (PTS) fin doping profile are susceptible to significant process-induced variations which result in lower manufacturing yield. Therefore, circuit-level "assist" techniques are commonly used to enhance the cell read margin and/or write margin to allow for lower minimum cell operating voltage (V_{MIN}) [5].

Inserted-oxide FinFET (iFinFET) technology was proposed to provide an evolutionary pathway for continued transistor scaling [6]. The electrostatic integrity of an iFinFET (*i.e.* gate control of the electrostatic potential in the channel region) is superior to that of a FinFET due to gate fringing electric fields through the inserted oxide (SiO_2) layers. Although the gate-all-around (GAA) field-effect transistor (FET) can achieve even better electrostatic integrity, this comes at the cost of a larger intrinsic delay [6]. Also, to achieve comparable

layout area efficiency as a FinFET, a GAAFET should comprise multiple nanowires (NWs) that are vertically stacked apart by 10 nm or more (to allow sufficient room for the dielectric/metal/dielectric gate stacks in-between the NWs), necessitating the formation of higher-aspect-ratio fin structures during the device fabrication process. In contrast, the inserted-oxide layers in an iFinFET can be very thin (less than 5 nm), and also can serve effectively as dopant diffusion barriers [7]. In this chapter, a scheme for controllably reducing the drive strength of an iFinFET is proposed and demonstrated via three-dimensional (3-D) device simulations and a calibrated compact model that to provide for lower V_{MIN} of a minimally sized 6-T SRAM cell, for high-density cache memory.

3.2 Inserted-Oxide FinFET (iFinFET)

It is generally expected that some form of gate-all-around transistor structure (GAAFET) will be used in the future to enable further reductions in gate length with adequate suppression of short-channel effects. To achieve similar or superior on-state current per unit layout area, the GAAFET must comprise multiple stacked nanowire (NW) channel regions. In practice, this is done by growing a "sandwich" of multiple $Si/Si_{1-x}Ge_x$ layers using multiple alternating epitaxy growth steps. The $Si_{1-x}Ge_x$ layers are later selectively etched away and replaced by the gate dielectric and gate metal layer stack. To accommodate these layers, the $Si_{1-x}Ge_x$ sacrificial layers must be relatively thick (~10nm). As a result, to form a GAAFET, a very high aspect ratio (height:width > 6:1) fin structure must be formed. To further improve the current density, more stacked NWs may be required, which means the aspect ratio will be even higher posing a greater fabrication challenge.

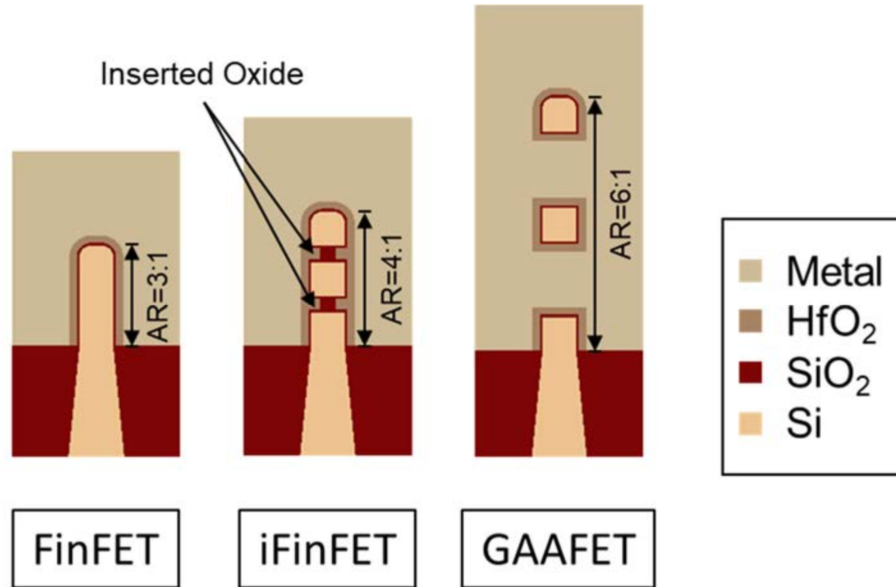


Figure 3.2-1. A comparison of the 3 transistor structures (Cross-sectional views across the channel regions). The fin widths are each 6nm, and the total Si height is 18nm. The inserted-oxide thickness is 3nm. The spacing between adjacent NWs in GAAFET is 10nm.

To mitigate these issues, an evolutionary FinFET design - the inserted-oxide FinFET (iFinFET) [6, 8, 9] is proposed (Figure 3.2-1). iFinFET exhibits superior electrostatic integrity by allowing the gate fringing field to penetrate the inserted-oxide and control the bottom of the NWs (except for the bottom NW), as shown in Figure 3.2-2. The performance can be further improved by recessing the inserted-oxide in the middle to replace portion of the inserted-oxide ($k = 3.9$) with HfO_2 ($k = 25$). This can be done after the dummy oxide removal and before the high-k dielectric (HfO_2) deposition in the replacement metal gate (RMG) module. Due to the small thickness (3nm) of the inserted-oxide, the gate metal will not be present in the recessed portion, and hence the increase in total capacitance is small. In addition, compared to the GAAFET, the aspect ratio of the fin structure that needs to be formed to make an iFinFET is much smaller.

The fabrication process of the iFinFET is identical to that of the conventional bulk-Si FinFET, except that a multi-SOI (silicon-on-insulator on silicon-on-insulator) wafer instead of a bulk-Si wafer is used as the starting substrate. As shown in [10], the 2-NW iFinFET also offers the benefit of superior electrostatic integrity over that of FinFET. In that case, a conventional SOI wafer can be used.

As shown in Figure 3.2-3, the subthreshold swing (SS_{SAT}) and drain-induced barrier lowering (DIBL) of the iFinFET is better than that of the FinFET, but worse than those of GAAFET. The situation is the same for the on-state currents and the intrinsic gain (g_m/g_{ds}). The gate-to-drain (C_{gd}) and total gate (C_{gg}) capacitance in iFinFET are larger than those of FinFET due to larger overlap between gate and the drain. This situation is exacerbated in the GAAFET. The n-channel iFinFET can achieve a similar intrinsic delay ($C_{gg}V_{DD}/I_{ON}$) as the FinFET. It should be noted that in this study an aggressively thin NW spacing of 6nm is assumed for the GAAFET. In practice this would be challenging to achieve.

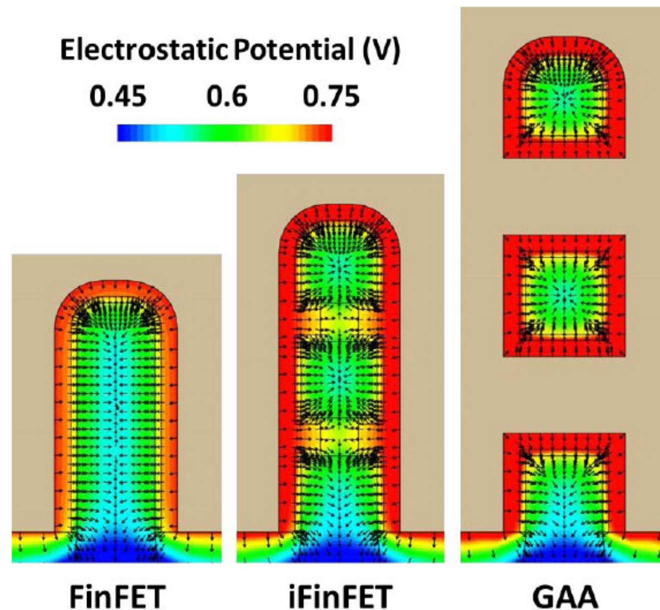


Figure 3.2-2. Electrostatic potential and electric field lines (arrows) in the linear regime (V_{GS}=0.75V, V_{DS}=0.05V) in the 3 transistor structures. It can be seen the iFinFET achieves good gate control over the electrostatic potential at the bottom of the NW (except for the bottom-most channel region). Adapted from [6].

With regard to process-induced variations, it has been shown that the performance of the iFinFET is relatively insensitive to variations in the thicknesses of the inserted-oxide and the inserted-oxide recess [6].

Device Parameter	FinFET / iFinFET / GAA
Nominal Gate Length, L_g (nm)	12
Fin/NW Width, W_{fin} (nm)	6
Total Si Height above STI (nm)	18
Gate Pitch (nm)	35
Equivalent Oxide Thickness (nm)	0.7
Inserted-Oxide Thickness (nm)	0 / 3 / 0
Metal Thickness between NWs (nm)	0 / 0 / 6
Specific Contact Resistivity (Ω-cm^2)	3.5×10^{-9}

	FinFET		iFinFET		GAA	
	N	P	N	P	N	P
$\|I_{OFF}\|$ (pA)	0.8	0.8	0.8	0.8	0.8	0.8
$\|I_{ON}\|$ (μA)	15.1	14.7	18.1	16.9	21.7	19.5
$\|I_{EFF}\|$ (μA)	7.1	7.1	8.6	8.2	10.4	9.5
$\|V_{T,SAT}\|$ (V)	0.42	0.42	0.41	0.40	0.40	0.39
DIBL (mV/V)	40	36	32	28	23	20
SS_{SAT} (mV/dec)	78	78	75	75	71	71
C_{gd} (aF)	8.1	7.8	9.2	8.9	12.0	11.6
C_{gg} (aF)	29.7	28.7	35.6	34.2	46.4	44.4
$C_{gg}V_{DD}/\|I_{ON}\|$ (ps)	1.47	1.45	1.47	1.52	1.60	1.70
g_m/g_{ds}	22.4	28.2	25.7	30.1	30.6	36.3

Figure 3.2-3. Design parameter values and simulated transistor performance values for the 3 different transistor structures (n-channel and p-channel). Adapted from [6].

3.3 6-T SRAM High Density Cell Design

3.3.1 6-T SRAM Operations

An SRAM circuit mainly consists of two parts: (1) the bit cells and (2) the peripheral circuits (Figure 3.3-1). The bit cells are arranged in rows and columns and each bit cell stores 1 bit of information. The peripheral circuits are circuits that surround the bit cells and are used to store and access information from the bit cells. Some specific functions of the peripheral circuits include (1) charging and discharging the bitlines (BLs and BLBs) and wordlines (WLs), (2) sensing the difference between BL and BLB potentials, (3) decoding inputs, etc. In

modern SRAM, complex peripheral circuits are also used to assist SRAM read and write operations [11, 12], at the cost of a larger layout area and greater power consumption.

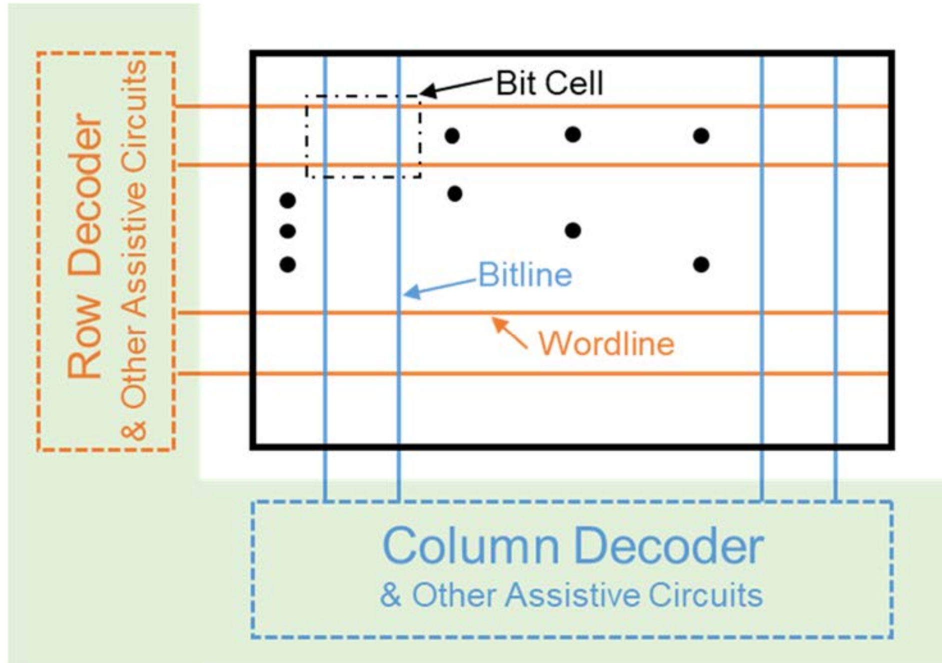


Figure 3.3-1. Block-level schematic of SRAM circuitry.

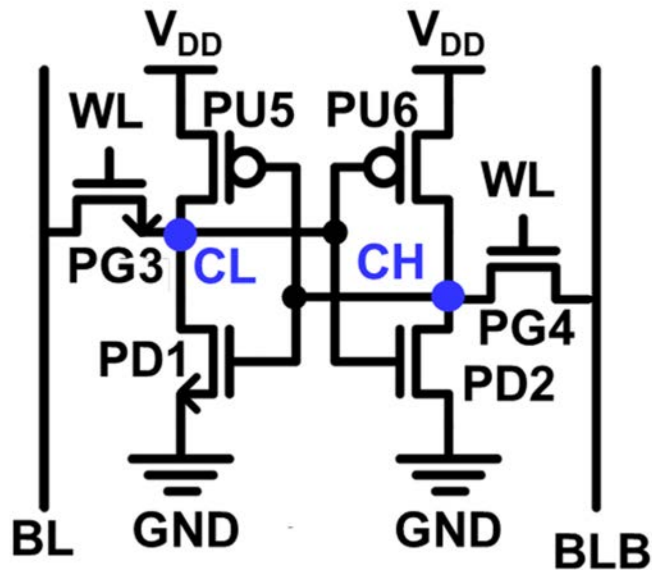


Figure 3.3-2. Circuit schematic of a 6-T SRAM bit cell.

In a 6-T SRAM, each bit cell comprises 6 transistors: (1) 2 pull-down (PD) n-channel and 2 pull-up (PU) p-channel transistors to form a pair of cross-coupled inverters, and (2) 2 pass-gate (PG) n-channel transistors to enable read and write access to the cross-coupled inverters. Figure 3.3-2 shows the schematic of a conventional 6-T SRAM bit cell. Information is stored at the output of the inverters (CL and CH).

In a read operation (Figure 3.3-3), the bitlines (BL and BLB) are precharged to the supply voltage V_{DD} by the peripheral circuits. Afterwards the wordline (WL) is pulsed to V_{DD} to turn on the PG transistors. Assuming CL stores "0" (0V) before this read operation, the charge stored in BL starts to flow through the conductive path from BL (at V_{DD}) and GND via PG3 (V_{GS}=V_{DD}-V_{CL}, V_{DS}=V_{DD}-V_{CL}) and PD1 (V_{GS}=V_{DD}, V_{DS}=V_{CL}). After some time, WL is discharged to 0V and PG transistors are turned off. By sensing the voltage difference between BL and BLB, the peripheral circuit determines the stored information, completing the read operation.

During the read process, V_{CL} is temporarily raised to some low voltage. In order to avoid changing the information stored (i.e., causing an accidental write operation), the aforementioned low voltage should not be large enough to turn on the opposite PD transistor (PD2 in Figure 3.3-3). Hence the drive strength of PD1 must be stronger than that of PG3. Note in this case, there is not significant current flowing through PG4 and PD2 since V_{DS}=0V for PG4 and V_{GS}=0V for PD2. Due to symmetry, PD2 must be stronger than PG4. By convention, the cell beta ratio is defined to be the ratio of the on-state current (drive strength) of a PD transistor to that of a PG transistor. For successful read operation, a large cell beta ratio is preferred.

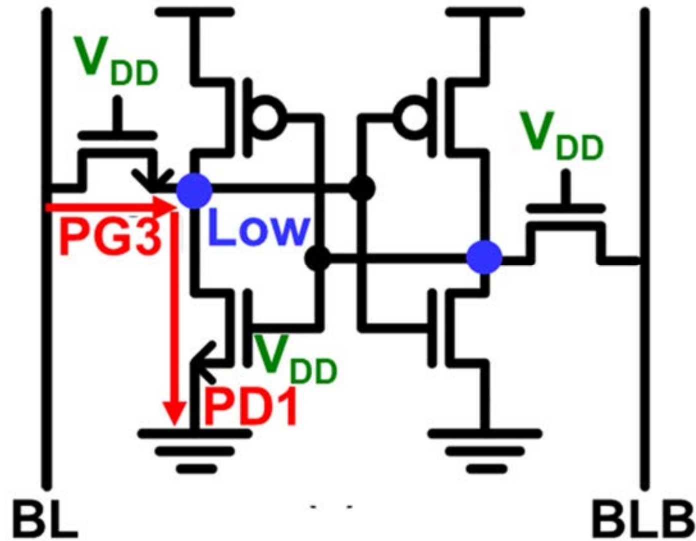


Figure 3.3-3. Read operation in a 6-T SRAM bit cell.

To quantify the robustness of a bit cell during a read operation, the read static noise margin (SNM) is used [13]. As shown in Figure 3.3-4, the SNM is defined as the minimum voltage noise applied to the internal node (CL or CH) to cause the stored information to flip in value. By plotting both V_{CL} vs. V_{CH} and V_{CH} vs. V_{CL} on the same plot to create the "butterfly" voltage transfer curves, the SNM can be extracted by the following method:

(1) Find the largest square that can fit in each of the two lobes of the butterfly plot. (If the two lobes are not perfectly symmetric, the smaller square is used for extraction.)

(2) SNM = the side length of this largest square.

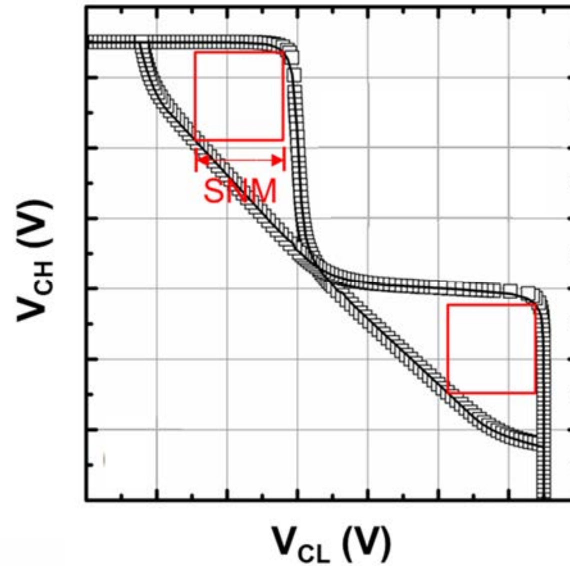


Figure 3.3-4. Definition of read SNM.

In a write operation, the bitlines (BL and BLB) are precharged to complementary logic values. Then the WL is pulsed to V_{DD} and the bitline at low voltage starts to discharge the corresponding internal node through the corresponding PG transistor. The voltage at the opposite internal node will eventually be raised to V_{DD} due to the nature of cross-coupled inverters. In Figure 3.3-5, V_{CH} is forced from V_{DD} to 0V by BLB. During this process, PU6 starts to conduct and tries to pull V_{CH} back to V_{DD}. Hence in order to ensure a successful write operation, PG4 should be stronger than PU6. Similarly, PG3 should be stronger than PU5. By convention, the cell gamma ratio is defined to be the ratio of the on-state current of PG transistor to that of PU transistor. For successful write operation, a large gamma ratio is preferred.

To quantify the write operation stability, the write N-curve is used [14]. The write N-curve can be obtained by sweeping the voltage at the internal node CL (CH) with BL (BLB) biased at V_{DD} and BLB (BL) biased at 0V, respectively, and measuring the current sourced into the internal node. The writability current (I_W) is defined to be the minimum current past the corresponding inverter (in this case, PD2 and PU6) tripping voltage (Figure 3.3-6). I_W must be positive to ensure a successful read. I_W is simply the difference between the current in PG and PU.

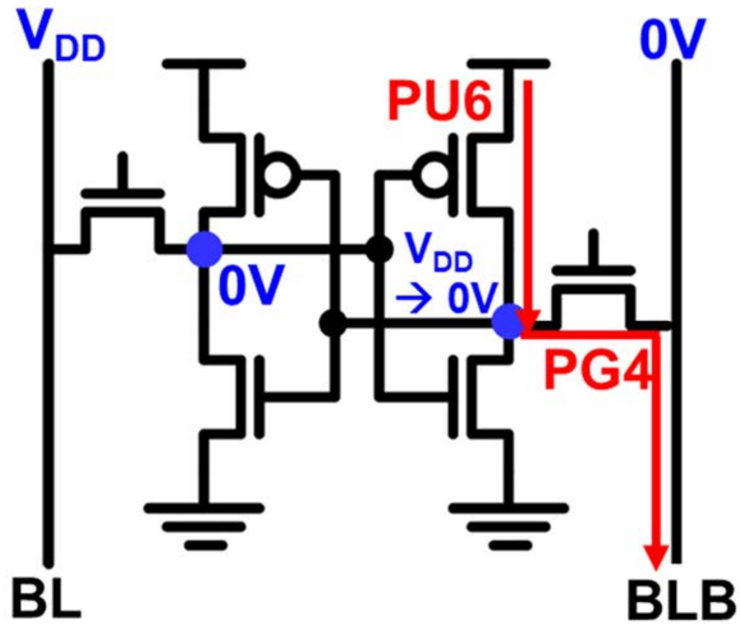


Figure 3.3-5. Write operation in a 6-T SRAM bit cell.

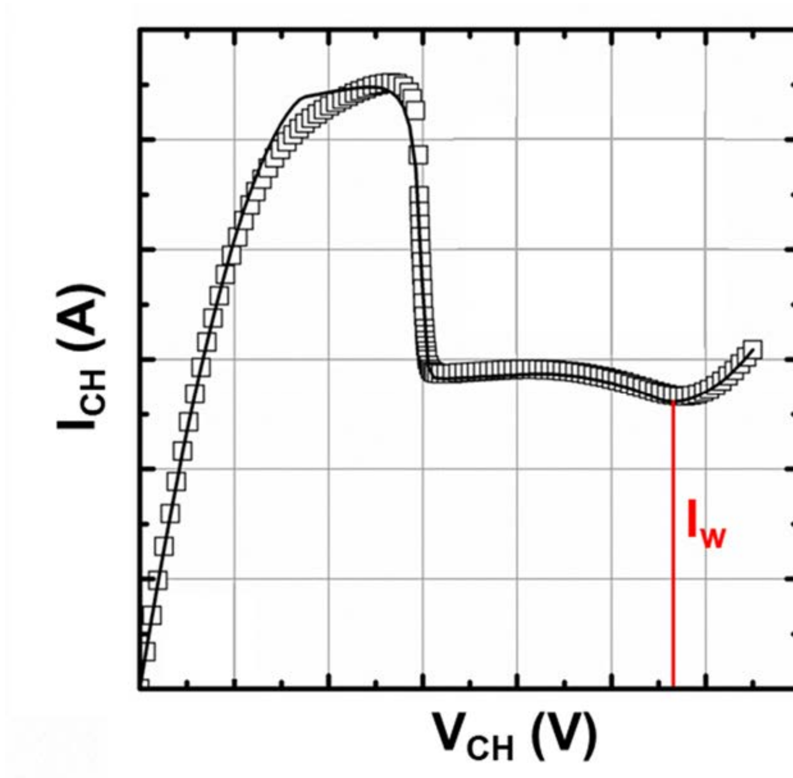


Figure 3.3-6. Definition of writability current.

3.3.2 FinFET High Density Cell Design

As was explained in Section 3.3.1, the PD n-channel transistor should be stronger than the PG n-channel transistor (to achieve a large cell beta ratio) for a successful read operation, and the PG n-channel transistor should be stronger than the PU p-channel transistor (to achieve a large cell gamma ratio). In planar FET technology, these two cell ratios are usually fine-tuned by changing the layout width of the transistors. However, this is not an option for FinFET technology because the effective channel width is quantized; SRAM designers can only adjust the discrete number of fins for each specific transistor. As a result, to achieve the most compact 6-T SRAM cell design (i.e., high density cell (HDC) design), all transistors must comprise 1 fin (Figure 3.3-7).

Due to the use of embedded SiGe source/drain stressors in p-channel FinFETs and {110}-oriented conduction surfaces in FinFETs, the hole mobility is much higher than in p-channel planar FETs. This results in comparable on-state current (per fin) between n-channel and p-channel FinFETs, making it challenging to meet the write stability requirement. In addition, since the (bulk) FinFET has a fully-depleted channel region, back biasing is not an effective means to tune the transistor threshold voltage (V_T). Consequently, peripheral assist circuitry is required to facilitate the operation of minimum-size FinFET 6-T SRAM cells.

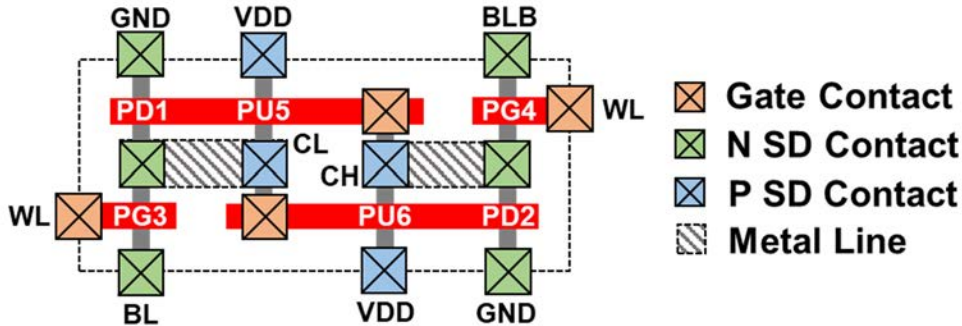


Figure 3.3-7. Sample FinFET HDC design layout.

A transistor-level solution would be preferable to save chip area and reduce power consumption. Previously proposed approaches [4, 15] include

1. selective STI recess so that PD transistors can have shallower STI and hence larger conduction width,
2. using longer gate length for PG and PU transistors, and
3. doping the fins in PG and PU transistors to fine tune their V_T's.

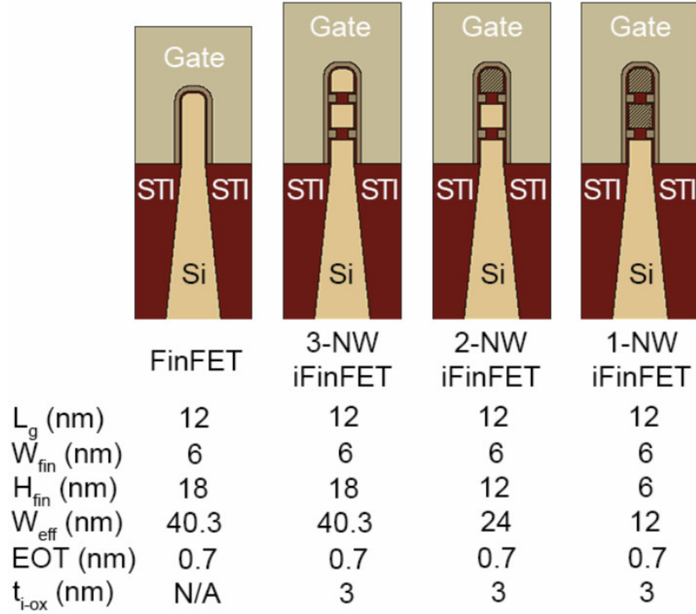
Each of the above approaches can result in additional process-induced variations. The last approach, in particular, is susceptible to random dopant fluctuation (RDF) effects.

3.4 High Density Cell Ratio Tuning Using Doped-Nanowire iFinFET

3.4.1 iFinFET Drive Strength Tuning

As discussed in Section 3.2, a 3-NW iFinFET can be fabricated using a conventional bulk-silicon FinFET process, starting with a silicon-on-insulator on silicon-on-insulator (SOI on SOI) substrate. In this section, the total silicon channel fin height is assumed to be the same for a 3-NW iFinFET as for the FinFET. Gate and channel dimensions are the same as in [6], appropriate for 4/3 nm CMOS technology. The effective channel width of the iFinFET can be controllably reduced by heavily doping the uppermost NW channel region(s) to render it (or them) non-conductive under normal transistor operating conditions.

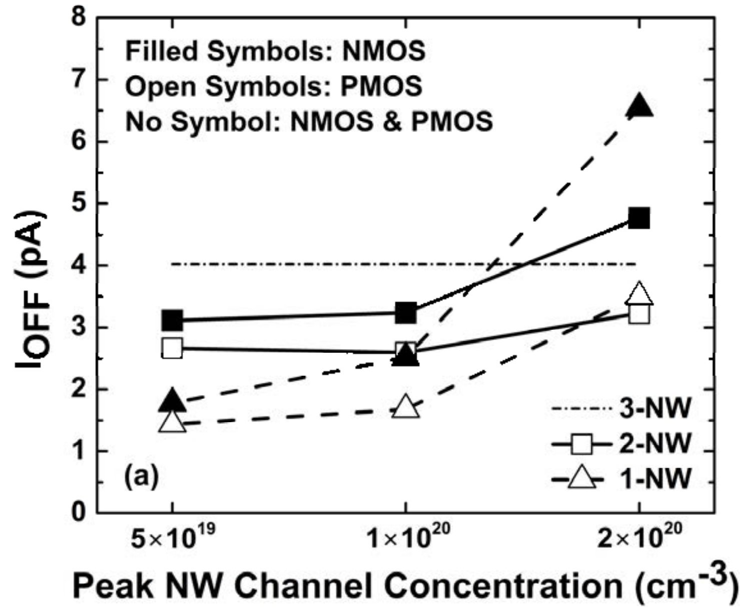
This doping can be achieved in practice by selective ion implantation during the STI formation module (See Section 3.6). Specifically, a photoresist mask is used to expose only the iFinFETs which are to receive the implant(s) that effectively increase the threshold voltage (V_T) of the uppermost NW(s) to be above the cell operating voltage (V_{DD}). This scheme can be used to selectively reduce the drive strengths of the PU iFinFETs and the PG iFinFETs (using a different mask and ion implantation step) in a 6-T SRAM cell. It should be noted that this scheme may be more difficult to implement for GAAFETs because of their larger NW separation which would necessitate deeper implant(s), *i.e.* larger projected range(s), which have larger straggle resulting in undesired doping of the lower NW channel region(s). Figure 3.4-1 shows the cross-section and structural parameters of the control (bulk-silicon) FinFET and three variants of the iFinFET.



	FinFET	3-NW iFinFET	2-NW iFinFET	1-NW iFinFET
L_g (nm)	12	12	12	12
W_{fin} (nm)	6	6	6	6
H_{fin} (nm)	18	18	12	6
W_{eff} (nm)	40.3	40.3	24	12
EOT (nm)	0.7	0.7	0.7	0.7
$t_{i\text{-}ox}$ (nm)	N/A	3	3	3

Figure 3.4-1. Cross-sectional views and channel dimensions of the simulated 3-D bulk-silicon FinFET and iFinFET structures in this work. The uppermost NW channel region(s) of the 2-NW and 1-NW iFinFETs are heavily doped, indicated by darker shading, so that they are non-conductive. The inserted oxide layers are slightly recessed due to dilute hydrofluoric acid treatment during the cleaning process prior to the formation of the high-k/metal gate stack. L_g is the nominal gate length. W_{fin} is the fin width. H_{fin} is defined to be the active Si channel height above STI. W_{eff} is defined as the outer perimeter of the active Si channel. EOT is the equivalent oxide thickness. $t_{i\text{-}ox}$ is the thickness of inserted oxide.

In contrast to the method of V_T tuning by ion implantation to adjust FinFET drive strength [15], the proposed scheme to increase V_T above V_{DD} for a subset of NW channels in an iFinFET should not result in significantly increased variation in transistor drive strength because the implanted channels do not contribute significantly to the transistor drive strength. The implanted NW channel dopant concentration should not be too high so as to result in large gate-induced drain leakage (GIDL). Thus, the criteria for optimizing the implanted NW channel dopant concentration are: (1) the off-state leakage current (I_{OFF}) should not be much higher than that of the nominal 3-NW iFinFET, and (2) the on-state drive current (I_{ON}) for the 2-NW (1-NW) iFinFET should be less than 2/3 (1/3) of that of a 3-NW iFinFET. This is because the top NW has larger effective channel width than the middle and bottom NWs, since its top surface is also gated.



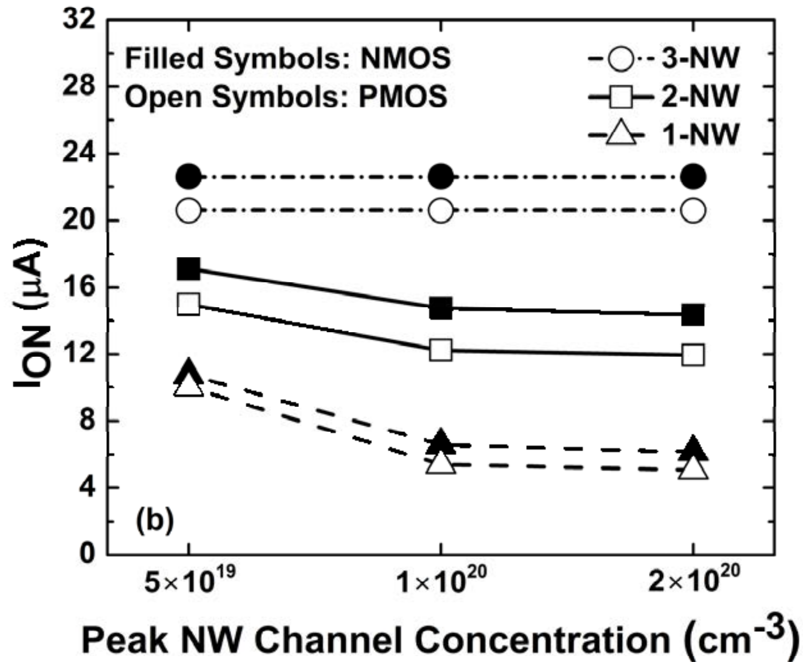


Figure 3.4-2. Effect of implanted NW channel dopant concentration on iFinFETs (a) off-state leakage current and (b) on-state drive current. Filled symbols correspond to NMOS iFinFETs; open symbols correspond to PMOS iFinFETs. I_{OFF} values are the same for 3-NW NMOS and PMOS iFinFETs.

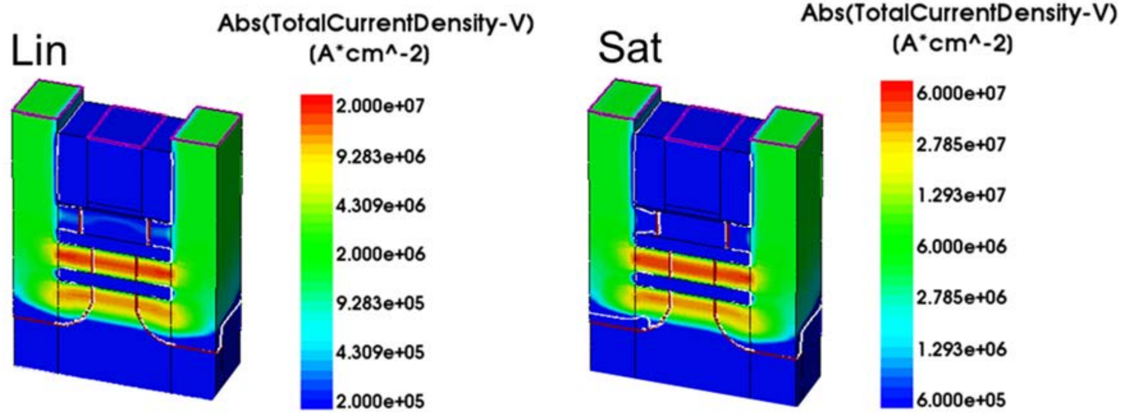


Figure 3.4-3. In a 2-NW iFinFET, the top (doped) NW does not conduct in on-state (Left: V_{GS}=0.75V, V_{DS}=0.05V, Right: V_{GS}=V_{DS}=0.75V). The doping concentration in the top NW is 1×10^{20}cm^{-3}.

3-D device simulations were performed using the technology computer aided design (TCAD) software tool Sentaurus Device [16] with calibrated models for carrier transport and quantum mechanical effects described in [6]. The gate work function, electrical channel length, and punchthrough-stopper doping profile were co-optimized to achieve maximum I_{ON} for I_{OFF} = 100pA/µm (normalized to effective channel width, W_{eff}) and V_{DD} = 0.75V for the nominal bulk FinFET and 3-NW iFinFET designs.

Figure 3.4-2(a) and (b) show how iFinFET I_{OFF} and I_{ON} values depend on the peak dopant concentration in the implanted NW channel(s), respectively. (Gaussian doping profiles with lateral steepness 3nm/dec are assumed.) From these plots it can be seen that 1×10^{20}cm^{-3} doping effectively reduces I_{ON} without dramatically increasing I_{OFF}; therefore, it is selected as the optimal dopant concentration for achieving a non-conducting NW channel. Table 3.4-1 summarizes key performance parameters of the control FinFET and iFinFETs. Figure 3.4-3 confirms that in a 2-NW iFinFET (top NW doped), the top NW does not conduct during on states.

		I_{OFF} (pA)	I_{ON} (µA)	$V_{TSAT}{}^{a}$ (V)	SS_{SAT} (mV/dec)
NMOS	3-NW	4.0	22.6	0.34	70
	2-NW	3.2	14.8	0.35	74
	1-NW	2.5	6.6	0.38	78
	FinFET	4.0	19.3	0.35	72
PMOS	3-NW	4.0	20.6	-0.34	70
	2-NW	2.6	12.2	-0.36	73
	1-NW	1.7	5.4	-0.39	78
	FinFET	4.0	18.8	-0.35	72

Table 3.4-1. Performance parameters for nominal transistor designs. (V_{DSAT}= V_{DD}=0.75V, V_{DLIN}=0.05V). ${}^{a}V_T$ is extracted at 100nA*W_{eff}/L_g.

3.4.2 High-Density 6-T SRAM Nominal Performance

To assess the nominal performance (read SNM and writeability current) of different HDC designs, a compact analytical MOSFET I-V model [17] is used herein to fit the transistor transfer characteristics simulated in TCAD (Section 3.4.1). As it can be seen from Figure 3.4-4, the compact model parameters are well calibrated to the TCAD results.

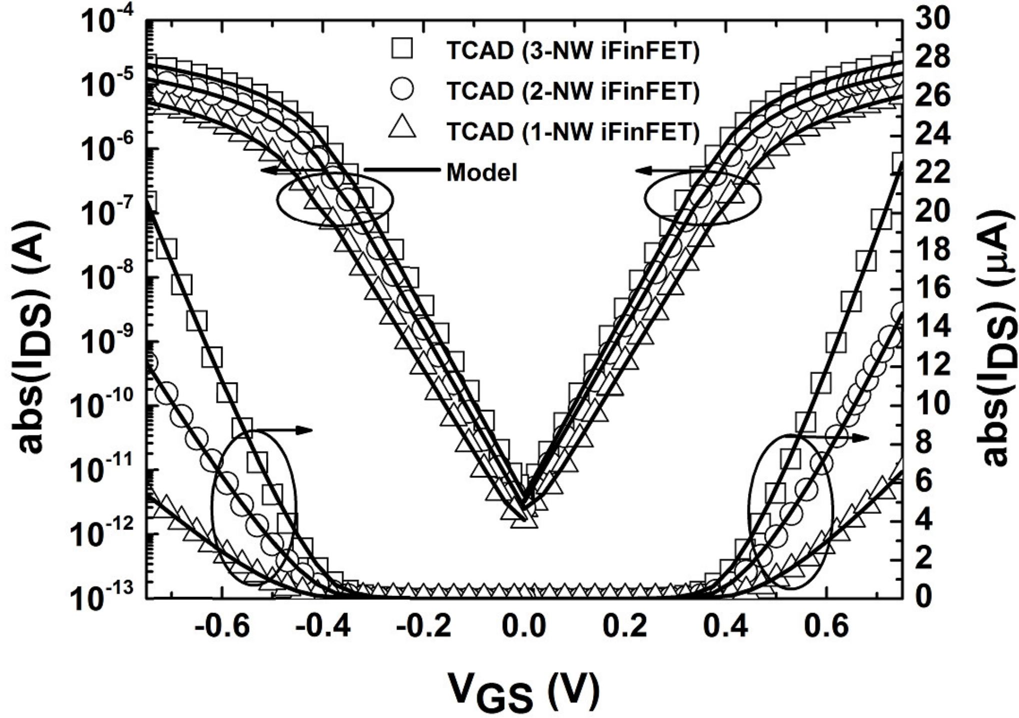


Figure 3.4-4. Comparison of simulated 1-NW, 2-NW and 3-NW iFinFET transfer characteristics with the fitted compact model, showing good agreement. (V_{DS}=V_{DD}=0.75V).

In the compact model, the inverter voltage transfer characteristics (VTC) curves are generated by solving Kirchhoff's current law (KCL) equations at the internal nodes (CL and CH) iteratively:

$I_{DS,PD1}(V_{GS}=V_{CH}, V_{DS}=V_{CL}) = I_{DS,PG3}(V_{GS}=V_{WL}-V_{CL}, V_{DS}=V_{BL}-V_{CL}) + I_{DS,PU5}(V_{GS}=V_{CH}-V_{DD}, V_{DS}=V_{CL}-V_{DD})$

$I_{DS,PD2}(V_{GS}=V_{CL}, V_{DS}=V_{CH}) = I_{DS,PG4}(V_{GS}=V_{WL}-V_{CH}, V_{DS}=V_{BLB}-V_{CH}) + I_{DS,PU6}(V_{GS}=V_{CL}-V_{DD}, V_{DS}=V_{CH}-V_{DD})$

Then the read SNM is extracted from the VTCs using the method mentioned in Section 3.3.1.

On the other hand, for write operation, the I_{CH} vs. V_{CH} and I_{CL} vs. V_{CL} curves are generated by calculating the values of I_{CH} and I_{CL} as V_{CH} and V_{CL} are swept:

$I_{CL} = I_{DS,PD1}(V_{GS}=V_{CH}, V_{DS}=V_{CL}) + I_{DS,PG3}(V_{GS}=V_{WL}-V_{CL}, V_{DS}=V_{BL}-V_{CL}) - I_{DS,PU5}(V_{GS}=V_{CH}-V_{DD}, V_{DS}=V_{CL}-V_{DD})$

$I_{CH} = I_{DS,PD2}(V_{GS}=V_{CL}, V_{DS}=V_{CH}) + I_{DS,PG4}(V_{GS}=V_{WL}-V_{CH}, V_{DS}=V_{BLB}-V_{CH}) - I_{DS,PU6}(V_{GS}=V_{CL}-V_{DD}, V_{DS}=V_{CH}-V_{DD})$

The writability current (I_W) is then extracted from these curves.

Figure 3.4-5 shows the inverter VTCs and I_{CH} vs. V_{CH} used for SNM and I_W extraction for the 2:3:3 iFinFET-based minimum-sized SRAM high desnsity cell design.

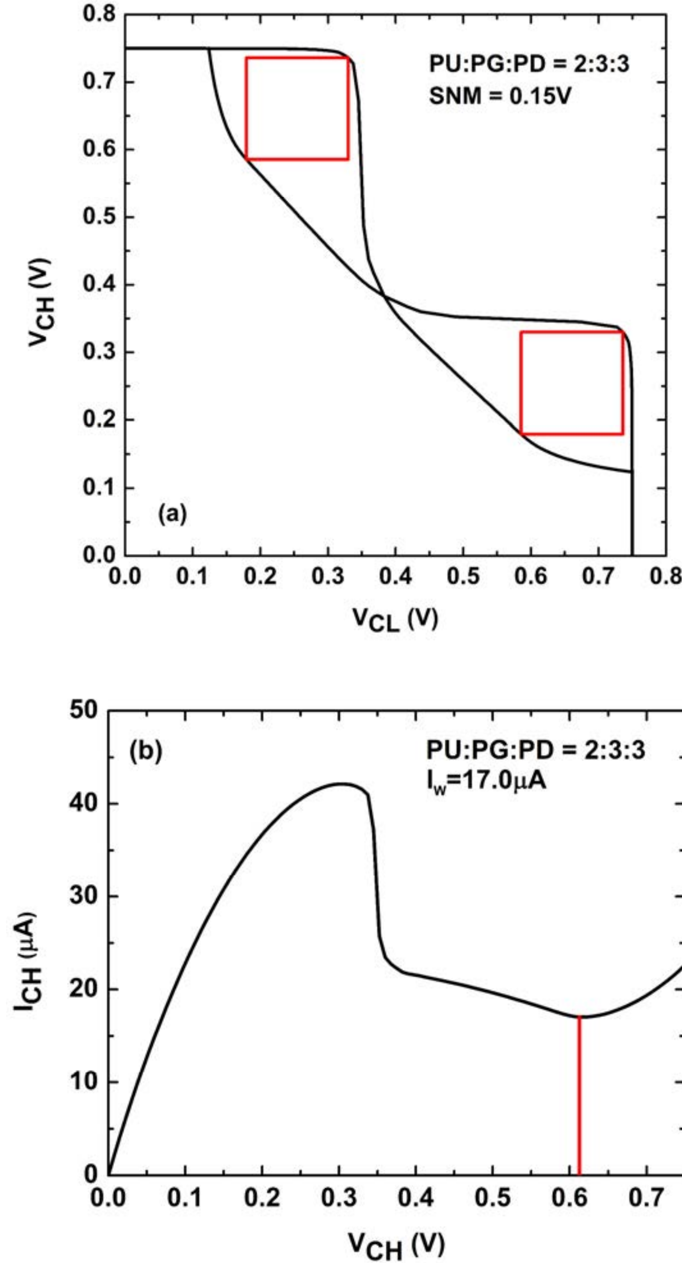


Figure 3.4-5. (a) Butterfly curve and (b) N-curve of the 2:3:3 design.

Design (PU: PG: PD NW ratio)	Nominal Cell Performance (V_{DD} = 0.75V)	
	SNM (V)	I_W (µA)
1:2:3	0.17	11.8
2:2:3	0.18	7.8
1:3:3	0.13	20.3
2:3:3	0.15	17.0
3:3:3	0.16	13.2
FinFET	0.15	10.7

Table 3.4-2. Nominal cell performance for different HDC designs.

For good read static noise margin (SNM) the PD devices should be the strongest; hence they are fixed to be 3-NW iFinFETs in this study. Table 3.4-2 summarizes the nominal performance of different HDC designs. By comparing 2:2:3 and 2:3:3 (also 1:2:3 and 1:3:3) designs, it can be seen the read SNM is improved by using a weaker PG. In addition, the read SNM is degraded when a weaker PU is used. For write operation, I_w is increased when PG is stronger than PU (2:3:3 vs. 3:3:3). Hence, to improve SNM the PG devices should be made weaker than the PD devices, whereas to improve write-ability current (I_W) the PU devices should be made weaker than the PG devices.

3.4.3 6-T SRAM HDC Yield Estimation

Transistor performance variability induced by process variations can result in failure in SRAM read (SNM < 0V) and write (I_W < 0A). We recognize two family of variation sources: (1) systematic and (2) random.

For systematic process-induced variations, we include variations in gate length (L_g) and fin width (W_{fin}). In this study, we assume they follow Gaussian distributions with mean value equal to their respective nominal values and 3 standard deviations equal to 10% of their nominal values. Specific to iFinFETs, it was shown in [6] that iFinFET performance is relatively insensitive to variations in the inserted-oxide thickness (t_{i-ox}) and inserted-oxide recess; hence these variations are neglected in this work (Figure 3.4-6).

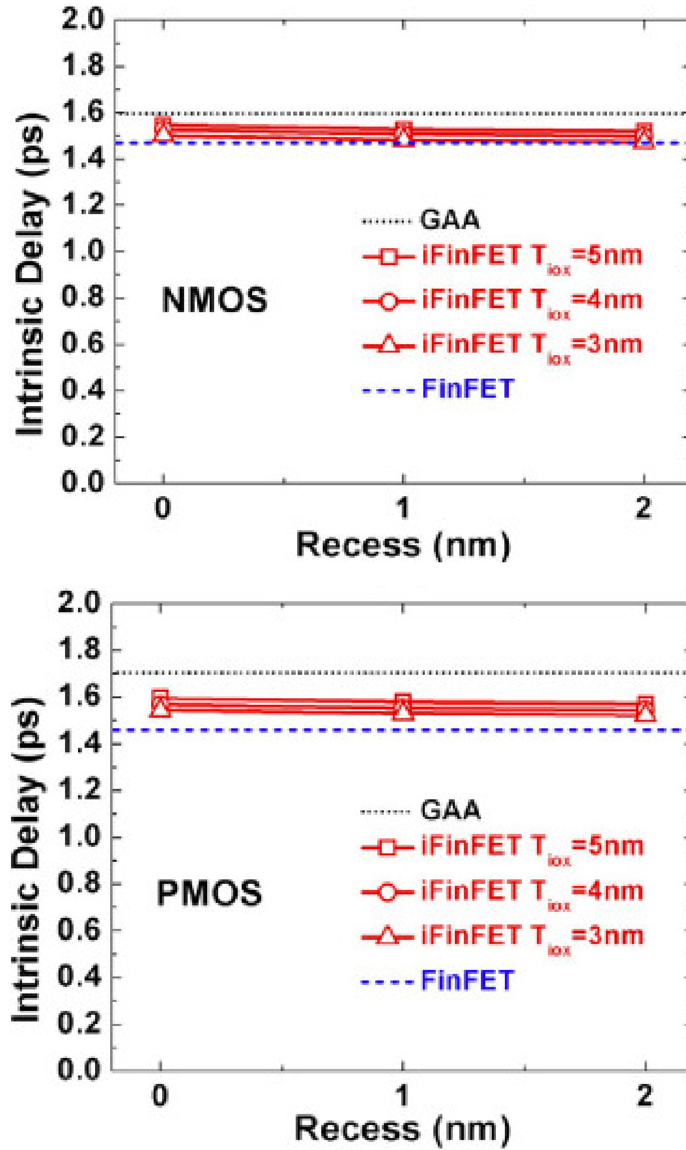


Figure 3.4-6. iFinFET performance is insensitive to the inserted-oxide thickness ($t_{i\text{-}ox}$). Adapted from [6].

We consider two random sources of variations: (1) random dopant fluctuation (RDF), and (2) gate workfunction variation (WFV). In sub-100nm planar FET technology, several ion implantation steps are used to fine-tune the V_T and reduce the transistor short-channel effects. V_T variation due to RDF in these planar FETs can be significant. This variation is generally due to the discrete nature of dopant atoms. As the transistor geometry continues to shrink, the actual number of dopant atoms within the transistor is very small. Due to the stochastic nature of ion implantation, having more or fewer dopant atoms in the transistor can cause large fluctuation in doping concentration [18]. As a

result, it is hard to precisely control V_T via fin doping in FinFET due to the tiny volume of the fin.

For FinFETs, previous research identified that WFV is the dominant contributor to V_T variation [19]. To suppress short channel effects in transistors with shorter gate length, the physical thickness of the gate oxide must be scaled down. This causes higher gate leakage due to reduced barrier width for carrier tunneling. To combat this problem, in 45nm technology node [20], a high-permittivity (high-k) gate dielectric (HfO_x) was introduced. The high-k dielectric material can achieve smaller electrical equivalent oxide thickness (EOT) with a larger physical thickness, hence suppressing the gate leakage. Due to the poor interface between poly-Si and high-k material, metal is selected as the gate material. The workfunction of the metal gate is determined by the sum of the metal's bulk chemical potential and the surface dipole potential. The latter property is dependent on the crystal orientation. Due to the stochastic nature of metal deposition, this value can vary and hence cause variation in the gate workfunction.

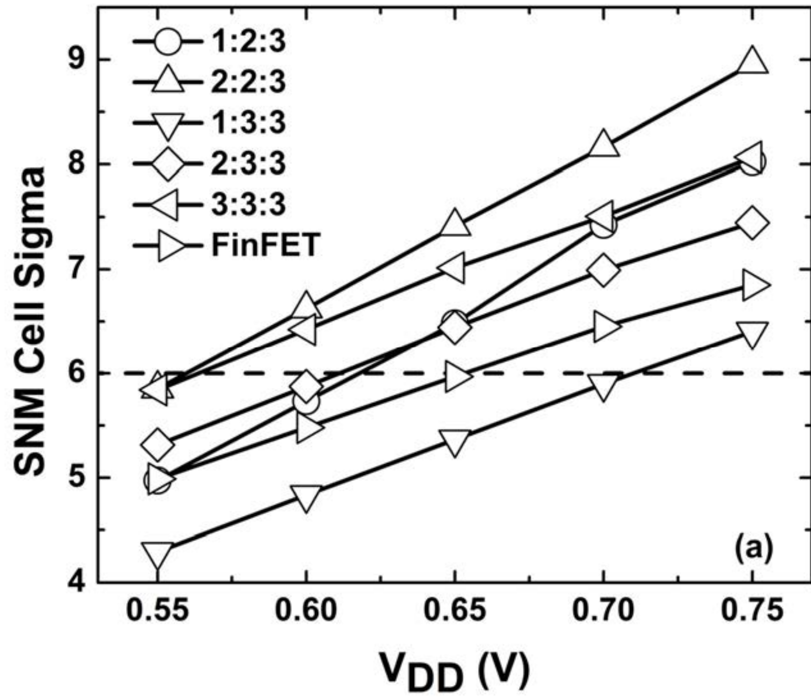
V_T variation due to random dopant fluctuations (RDF) and gate work function variation (WFV) [21, 22, 23] were simulated using the noise-like Impedance Field Method (n-IFM) in Sentaurus Device [16].

Table 3.4-3 compares the values of σ_{VT} due to random sources, for different transistor designs. WFV dominates random variation in V_T, and becomes worse as the number of NW channels is reduced [22]. The effect of RDF also increases with decreasing total channel volume, as expected. In this work, WFV and RDF are assumed to be independent and the total variation in V_T due to random sources are calculated by:

$$\sigma_{VT,total} = \sqrt{\sigma_{VT,RDF}^2 + \sigma_{VT,WFV}^2}$$

		σ_{VTSAT}					
		RDF Only		WFV Only		RDF & WFV	
		Abs. (mV)	Norm.[a]	Abs. (mV)	Norm.	Abs. (mV)	Norm.
NMOS	3-NW	2.30	0.007	25.31	0.073	25.41	0.074
	2-NW	5.01	0.014	29.58	0.084	30.00	0.085
	1-NW	11.08	0.029	36.54	0.095	38.19	0.099
	FinFET	2.92	0.008	29.39	0.083	29.53	0.083
PMOS	3-NW	3.84	0.011	25.13	0.073	25.42	0.074
	2-NW	6.63	0.019	29.35	0.082	30.09	0.084
	1-NW	13.55	0.035	36.02	0.092	38.49	0.098
	FinFET	4.33	0.012	28.64	0.081	28.97	0.082

Table 3.4-3. V_T variation due to random sources. [a]Normalized values are calculated by dividing σ_{VTSAT} by the corresponding nominal V_{TSAT} (listed in Table 3.4-1).



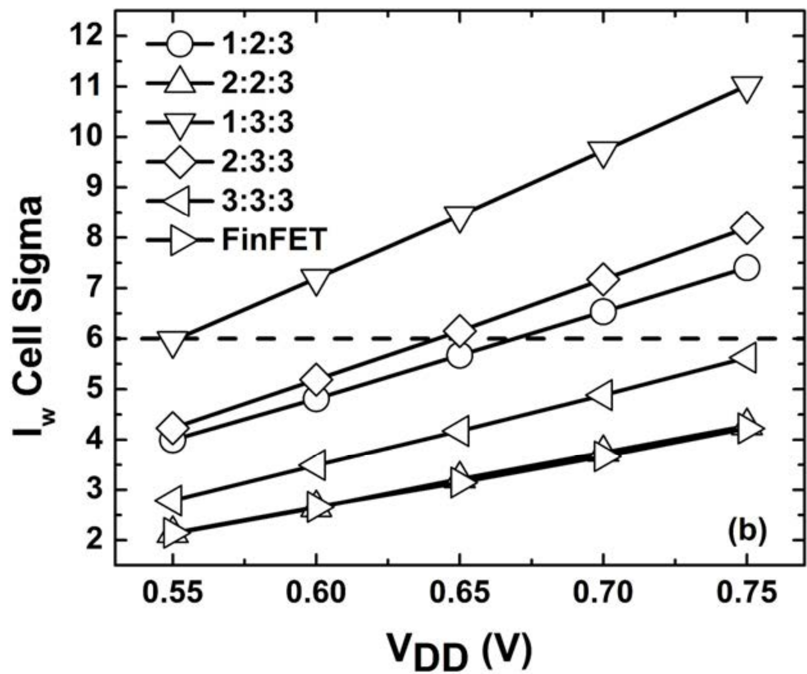


Figure 3.4-7. (a) SNM cell sigma vs. V_{DD} and (b) I_W cell sigma vs. V_{DD} for various 6-T SRAM high density cell designs. Cell sigma is defined to be the smaller one of SNM cell sigma and I_W cell sigma. The dashed line shows the 6 sigma reference line.

Design **(PU: PG: PD NW ratio)**	**V_{MIN}[a] (V)** **(Read / Write)**
1:2:3	0.62 / **0.67**
2:2:3	0.56 / >**0.75**
1:3:3	**0.71** / 0.55
2:3:3	0.61 / **0.64**
3:3:3	0.56 / >**0.75**
FinFET	0.65 / >**0.75**

Table 3.4-4. 6-T SRAM high-density yield estimation. [a]For a cell design with $V_{MIN}>V_{DD}=0.75V$, read/write assist techniques are required. The voltage values highlighted in bold-face font set V_{MIN}.

The manufacturing yield of the SRAM is quantified by the cell sigma. The cell sigma is defined as the minimum total number of standard variations from the respective nominal values in all 3 sources of variation (i.e., L_g variation, W_{fin} variation, V_T variation due to RDF and WFV) for all the 6 transistors in the cell to cause a negative read SNM (read failure) or a negative I_W (write failure). Within each transistor, all variation sources are assumed to be independent. In addition, variations in different transistors are also assumed to be independent. Hence in the compact model, the variation space is modeled as 18-dimensional (6 transistors × 3 variation sources/transistor). The origin corresponds to the nominal device with no variations. Hence there exists a (hyper-)surface of failures, which are combinations of variations that result in read or write failures. The cell sigma is then calculated by finding the shortest distance from this surface of failures to the origin, which is assumed to be the most likely failure case.

Table 3.4-4 summarizes the yield of various 6-T SRAM HDC designs. Figure 3.4-7(a) and (b) plot SNM and I_W cell sigmas as a function of V_{DD}. The 2:3:3 (PU:PG:PD NW ratio) cell design achieves the lowest V_{MIN} (0.64V) due to a good balance between read and write margins with minimal increase in process-induced variations. This V_{MIN} value is comparable to that for a 14 nm-generation FinFET-based high-density SRAM cell design (without assist) [24]. Thus our proposed scheme facilitates SRAM cell area scaling without a trade-off in V_{MIN}.

To implement the optimal (2:3:3) cell design, only one extra lithography mask is needed.

3.5 Conclusion

In this chapter, a novel scheme for controllably reducing the drive strength of a 3-D transistor is proposed and shown through simulations to facilitate voltage scaling of a minimally sized 6-T SRAM cell. Specifically, one or more of the stacked nanowire channels within an iFinFET can be made to be essentially non-conducting by ion implantation to increase its threshold voltage. Via three-dimensional device simulations and a calibrated compact I-V model, this scheme is projected to enable more than 0.1V reduction in minimum cell operating voltage (V_{MIN}) for a 6-T SRAM high-density cell design. This technique can also be applied to other cell designs, e.g., employing multiple-fin devices, for higher-speed and/or lower-power operation.

3.6 Appendix: Doped-Nanowire iFinFET Fabrication Process Issues

3.6.1 Feasibility of Using Ion Implantation to Dope Top NW(s) in iFinFET

In a doped-nanowire iFinFET, the top NW(s) are doped using ion implantation. In order not to accidentally contaminate the NW(s) below, the conditions of ion implantation must be carefully selected. In addition, after the implant step, a high-temperature diffusion process must be used to activate the dopants in these doped NW(s). Hence it is critical that the inserted-oxide layers between Si NWs can serve as dopant diffusion barriers so that the dopants in the upper NWs do not diffuse into lower NWs.

To validate the feasibility of this implant with the aforementioned constraint, the Synopsys process simulator, S-Process [16], was used to simulate the implant profile as implanted and after diffusion. The advanced calibration process models in S-Process were turned on to obtain the most accurate ion implant and diffusion result. It is shown in Section 3.4 that the best design features a 2-NW p-channel iFinFET for PU and 3-NW n-channel iFinFET for PD and PG. Hence

only p-channel iFinFET needs to be doped by donors. To control the straggle in the implant, the heavier elemental dopant Arsenic is selected. Figure 3.6-1 shows the implant profiles as implanted and after diffusion in two cases: (1) implant without a screening oxide layer on top and (2) implant through a 5nm screening oxide layer. The screening oxide in case (2) is stripped immediately after the implant and prior to thermal annealing. The boron background profile is included as reference. Both cases are subjected to a 1000°C 30min annealing right after the implant.

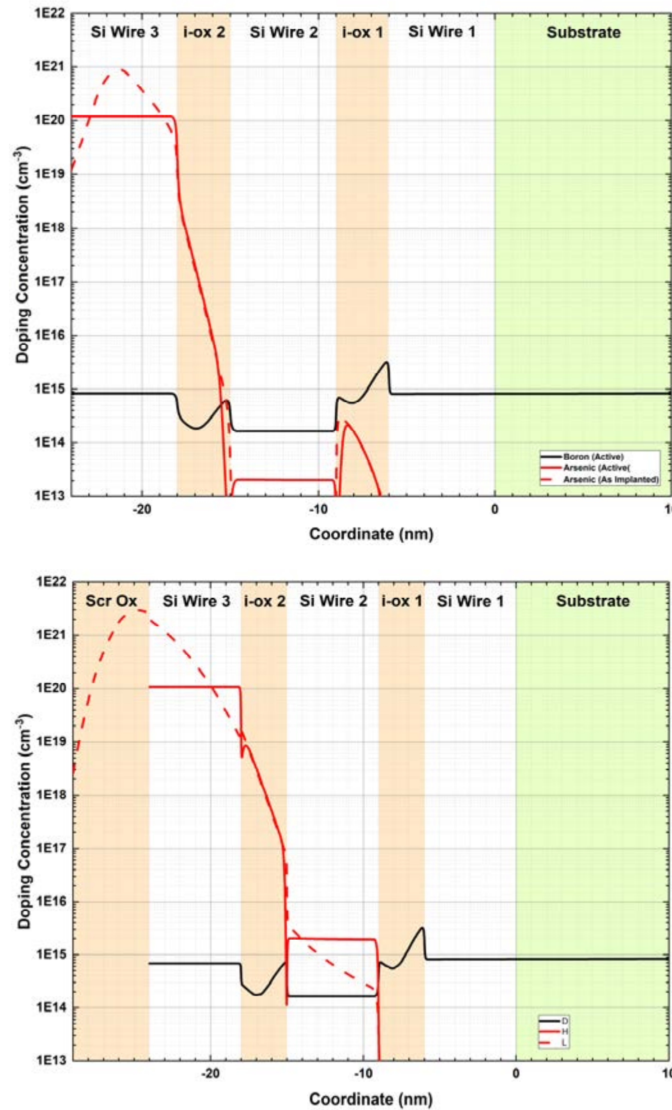


Figure 3.6-1. Implant profiles in p-channel 2-NW iFinFETs (only the top NW is doped). Top: Implant without a screening oxide. Bottom: Implant through a 5nm screening oxide. The horizontal axis shows the coordinates along the height dimension. The substrate is not shown to the full scale.

	T_{scox} (nm)	Species	Energy (keV)	Tilt (°)	Dose (cm^{-2})
Case (1)	0	Arsenic	1	7	2.5×10^{14}
Case (2)	5	Arsenic	2	7	1×10^{15}

Table 3.6-1. Implant conditions for the two cases: (1) implant without a screening oxide, and (2) implant through a 5nm screening oxide.

From Table 3.6-1, it can be seen that when implanting without a screening oxide, the first NW can achieve a $1\times10^{20}cm^{-3}$ Arsenic doping in the top NW, while the Arsenic concentrations in the middle NW and bottom NW are < $1\times10^{15}cm^{-3}$, smaller than the Boron background doping. Implanting through a 5nm screening oxide requires a higher dose (due to dopant loss when stripping the screening oxide) and a higher energy to meet $1\times10^{20}cm^{-3}$ doping concentration in the top NW. The implant straggle is larger than that in case (1). However, in this case, the Arsenic concentration in the middle NW is still < $1\times10^{15}cm^{-3}$. The change in transistor characteristics is negligible. Therefore, we can conclude the ion implantation in either case does not affect the bottom two NWs.

3.6.2 Proposed Doped-Nanowire iFinFET SRAM Fabrication Process

As shown in Section 3.4, the best design (PU:PG:PD=2:3:3) only requires the PU p-channel iFinFET to be doped. One approach to do this is to add a masked ion implantation during the STI formation module. Figure 3.6-2 shows the proposed process flow for this best iFinFET HDC design. It should be noted that the dopant activation step (step (f)) can be combined with later thermal annealing process to avoid excessive dopant diffusion.

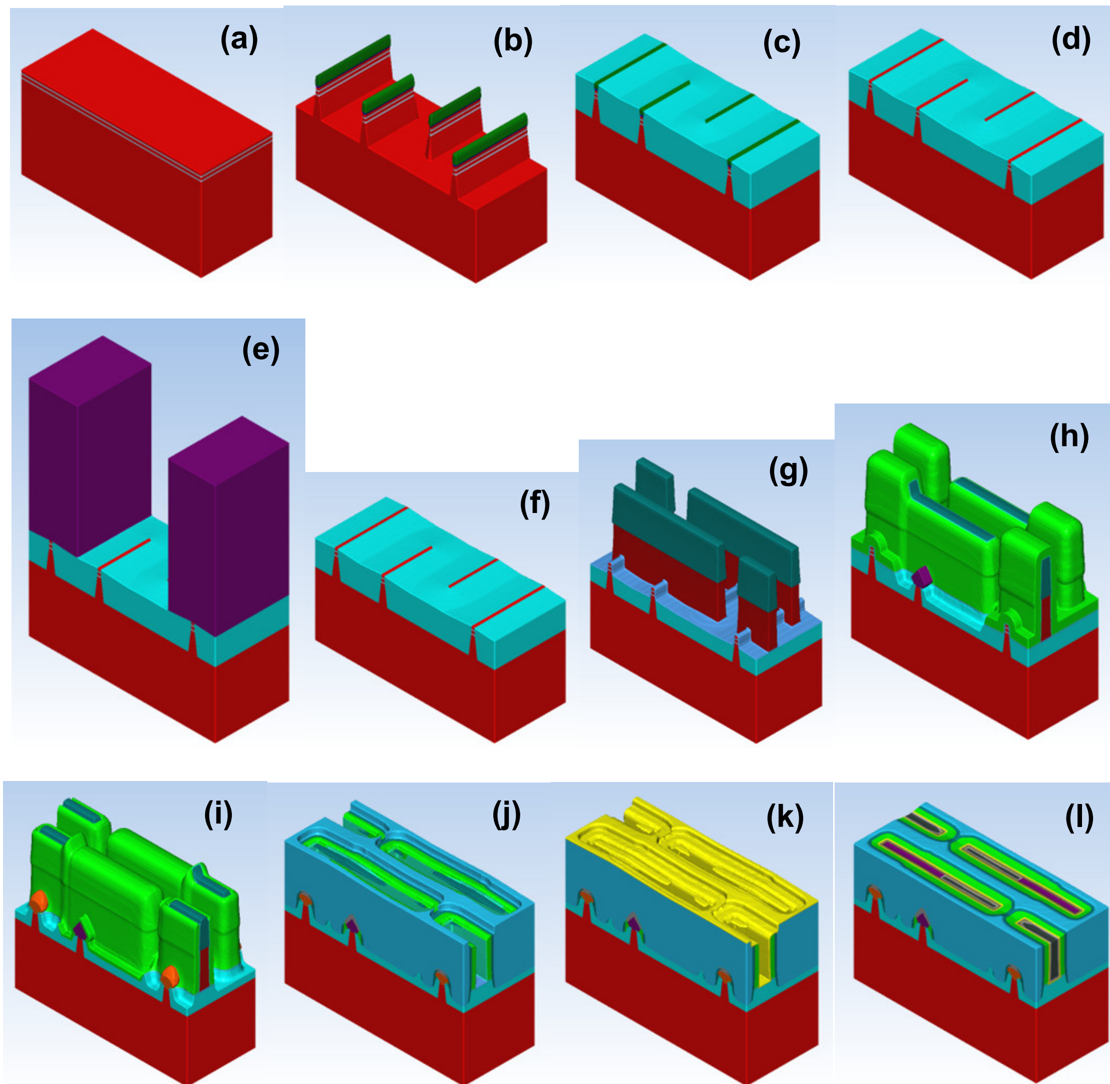


Figure 3.6-2. Proposed process flow for the best doped-nanowire iFinFET HDC SRAM design (PU:PG:PD=2:3:3). Only selected steps are shown: (a) starting SOI-on-SOI substrate as required by a 3-NW iFinFET; (b) fin patterning; (c) STI oxide fill and CMP (stopping material: Nitride); (d) recess the nitride hard mask to expose the top wire of iFinFET; (e) selective ion implantation in PMOS iFinFETs; (f) 1000°C 30min annealing/diffusion; (g) dummy gate patterning; (h) SiGe epitaxy growth (PMOS only); (i) SiP epitaxy growth (NMOS only); (j) dummy gate removal. (k) high-k dielectric deposition; (l) P&N workfunction metal layers deposition and W plug deposition. The middle-of-line (MOL) and backend-of-line (BEOL) are not shown.

Figure 3.6-3. Proposed additional mask for selectively doping the top wire of PU PMOS iFinFET.

In this particular case, only 1 additional mask is required (Figure 3.6-3). The process flow of 6-T SRAM HDC is emulated using Coventor SEmulator3D [25]. It is also possible that this selective ion implantation is done right after the dummy gate removal in the replacement metal gate integrated process module.

3.7 References

[1] S. Natarajan, M. Agostinelli, S. Akbar, M. Bost, A. Bowonder, V. Chikarmane, S. Chouksey, A. Dasgupta, K. Fischer, Q. Fu, T. Ghani, M. Giles, S. Govindaraju, R. Grover, W. Han, D. Hanken, E. Haralson, M. Haran, M. Heckscher, R. Heussner, P. Jain, R. James, R. Jhaveri, I. Jin, H. Kam, E. Karl, C. Kenyon, M. Liu, Y. Luo, R. Mehandru, S. Morarka, L. Neiberg, P. Packan, A. Paliwal, C. Parker, P. Patel, R. Patel, C. Pelto, L. Pipes, P. Plekhanov, M. Prince, S. Rajamani, J. Sandford, B. Sell, S. Sivakumar, P. Smith, B. Song, K. Tone, T. Troeger, J. Wiedemer, M. Yang and K. Zhang, "A 14nm logic technology featuring 2nd-generation FinFET, air-gapped interconnects, self-aligned double patterning and a 0.0588 µm2 SRAM cell size," 2014 IEEE International Electron Devices Meeting, San Francisco, CA, 2014, pp. 3.7.1-3.7.3. doi: 10.1109/IEDM.2014.7046976.

[2] S. Y. Wu, C. Y. Lin, M. C. Chiang, J. J. Liaw, J. Y. Cheng, S. H. Yang, S. Z. Chang, M. Liang, T. Miyashita, C. H. Tsai, C. H. Chang, V. S. Chang, Y. K. Wu, J. H. Chen, H. F. Chen, S. Y. Chang, K. H. Pan, R. F. Tsui, C. H. Yao, K.

C. Ting, T. Yamamoto, H. T. Huang, T. L. Lee, C. H. Lee, W. Chang, H. M. Lee, C. C. Chen, T. Chang, R. Chen, Y. H. Chiu, M. H. Tsai, S. M. Jang, K. S. Chen and Y. Ku, "An enhanced 16nm CMOS technology featuring 2nd generation FinFET transistors and advanced Cu/low-k interconnect for low power and high performance applications," 2014 IEEE International Electron Devices Meeting, San Francisco, CA, 2014, pp. 3.1.1-3.1.4. doi: 10.1109/IEDM.2014.7046970.

[3] D. Burnett, S. Parihar, H. Ramamurthy and S. Balasubramanian, "FinFET SRAM design challenges," 2014 IEEE International Conference on IC Design & Technology, Austin, TX, 2014, pp. 1-4. doi: 10.1109/ICICDT.2014.6838606.

[4] H. Kawasaki, K. Okano, A. Kaneko, A. Yagishita, T. Izumida, T. Kanemura, K. Kasai, T. Ishida, T. Sasaki, Y. Takeyama, N. Aoki, N. Ohtsuka, K. Suguro, K. Eguchi, Y. Tsunashima, S. Inaba, K. Ishimaru and H. Ishiuchi, "Embedded Bulk FinFET SRAM Cell Technology with Planar FET Peripheral Circuit for hp32 nm Node and Beyond," 2006 Symposium on VLSI Technology, 2006. Digest of Technical Papers., Honolulu, HI, 2006, pp. 70-71. doi: 10.1109/VLSIT.2006.1705221.

[5] E. Karl, Z. Guo, J. W. Conary, J. L. Miller, Y. G. Ng, S. Nalam, D. Kim, J. Keane, U. Bhattacharya and K. Zhang, "17.1 A 0.6V 1.5GHz 84Mb SRAM design in 14nm FinFET CMOS technology," 2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers, San Francisco, CA, 2015, pp. 1-3. doi: 10.1109/ISSCC.2015.7063050.

[6] P. Zheng, D. Connelly, F. Ding and T. J. K. Liu, "FinFET Evolution Toward Stacked-Nanowire FET for CMOS Technology Scaling," in IEEE Transactions on Electron Devices, vol. 62, no. 12, pp. 3945-3950, Dec. 2015. doi: 10.1109/TED.2015.2487367.

[7] J. D. Plummer, M. D. Deal and P. B. Griffin, Silicon VLSI Technology: Fundamentals, Practice and Modeling. Upper Saddle River, NJ: Prentice Hall, 2000.

[8] P. Zheng, D. Connelly, F. Ding and T. K. Liu, "Inserted-oxide FinFET (iFinFET) design to extend CMOS scaling," 2015 International Symposium on VLSI Technology, Systems and Applications, Hsinchu, 2015, pp. 1-2. doi: 10.1109/VLSI-TSA.2015.7117573.

[9] P. Zheng, D. Connelly, F. Ding and T. K. Liu, "Simulation-Based Study of the Inserted-Oxide FinFET for Future Low-Power System-on-Chip Applications," in IEEE Electron Device Letters, vol. 36, no. 8, pp. 742-744, Aug. 2015. doi: 10.1109/LED.2015.2438856.

[10] Y. Wu, F. Ding, D. Connelly, M. Chiang, J. F. Chen and T. K. Liu, "Simulation-Based Study of High-Density SRAM Voltage Scaling Enabled by Inserted-Oxide FinFET Technology," in IEEE Transactions on Electron Devices, vol. 66, no. 4, pp. 1754-1759, April 2019. doi: 10.1109/TED.2019.2900921.

[11] B. Zimmer, S. O. Toh, H. Vo, Y. Lee, O. Thomas, K. Asanovic and B. Nikolic, "SRAM Assist Techniques for Operation in a Wide Voltage Range in 28-nm CMOS," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 59, no. 12, pp. 853-857, Dec. 2012. doi: 10.1109/TCSII.2012.2231015.

[12] M. F. Chang, C. F. Chen, T. H. Chang, C. C. Shuai, Y. Y. Wang, Y. J. Chen and H. Yamauchi, "A Compact-Area Low-VDDmin 6T SRAM With Improvement in Cell Stability, Read Speed, and Write Margin Using a Dual-Split-Control-Assist Scheme," in IEEE Journal of Solid-State Circuits, vol. 52, no. 9, pp. 2498-2514, Sept. 2017. doi: 10.1109/JSSC.2017.2701547.

[13] E. Seevinck, F. J. List and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," in IEEE Journal of Solid-State Circuits, vol. 22, no. 5, pp. 748-754, Oct. 1987. doi: 10.1109/JSSC.1987.1052809.

[14] C. Wann, R. Wong, D.J. Frank, R. Mann, S.-B. Ko, P. Croce, D. Lea, D. Hoyniak, Y.-M. Lee, J. Toomey, M. Weybright and J. Sudijono, "SRAM cell design for stability methodology," IEEE VLSI-TSA International Symposium on VLSI Technology, 2005. (VLSI-TSA-Tech)., Hsinchu, 2005, pp. 21-22. doi: 10.1109/VTSA.2005.1497065.

[15] C.-H. Lin, K. K. Das, L. Chang, R. Q. Williams, W. E. Haensch and C. Hu, "VDD Scaling for FinFET Logic and Memory Circuits: the Impact of Process Variations and SRAM Stability," 2006 International Symposium on VLSI Technology, Systems, and Applications, Hsinchu, 2006, pp. 1-2. doi: 10.1109/VTSA.2006.251056.

[16] Sentaurus User's Manual, Version L-2016.03, Synopsys, Inc., Mountain View, CA, USA.

[17] A. E. Carlson, “Device and circuit techniques for reducing variation in nanoscale SRAM,” Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Univ. California Berkeley, Berkeley, CA, USA, 2008.

[18] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 /spl mu/m MOSFET's: A 3-D "atomistic" simulation study," in IEEE Transactions on Electron Devices, vol. 45, no. 12, pp. 2505-2513, Dec. 1998. doi: 10.1109/16.735728.

[19] T. Matsukawa, S. O'uchi, K. Endo, Y. Ishikawa, H. Yamauchi, Y.X. Liu, J. Tsukada, K. Sakamoto and M. Masahara, "Comprehensive analysis of variability sources of FinFET characteristics," 2009 Symposium on VLSI Technology, Honolulu, HI, 2009, pp. 118-119.

[20] 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 and C. Wiegand, "45nm High-k + metal gate strain-enhanced transistors," 2008 Symposium on VLSI Technology, Honolulu, HI, 2008, pp. 128-129. doi: 10.1109/VLSIT.2008.4588589.

[21] X. Wang, A. R. Brown, B. Cheng and A. Asenov, "Statistical variability and reliability in nanoscale FinFETs," Electron Devices Meeting (IEDM), 2011 IEEE International, Washington, DC, 2011, pp. 5.4.1-5.4.4. doi: 10.1109/IEDM.2011.6131494.

[22] H. F. Dadgour, K. Endo, V. K. De and K. Banerjee, "Grain-Orientation Induced Work Function Variation in Nanoscale Metal-Gate Transistors—Part I: Modeling, Analysis, and Experimental Validation," in IEEE Transactions on Electron Devices, vol. 57, no. 10, pp. 2504-2514, Oct. 2010. doi: 10.1109/TED.2010.2063191.

[23] X. Zhang, J. Li, M. Grubbs, M. Deal, B. Magyari-Köpe, B. M. Clemens and Y. Nishi, "Physical model of the impact of metal grain work function variability on emerging dual metal gate MOSFETs and its implication for SRAM reliability," 2009 IEEE International Electron Devices Meeting (IEDM), Baltimore, MD, 2009, pp. 1-4. doi: 10.1109/IEDM.2009.5424420.

[24] T. Song, W. Rim, J. Jung, G. Yang, J. Park, S. Park, Y. Kim, K.-H. Baek, S. Baek, S.-K. Oh, J. Jung, S. Kim, G. Kim, J. Kim, Y. Lee, S.-P. Sim, J. S. Yoon, K.-M. Choi, H. Won and J. Park, "A 14 nm FinFET 128 Mb SRAM With VMIN Enhancement Techniques for Low-Power Applications," in IEEE Journal of Solid-State Circuits, vol. 50, no. 1, pp. 158-169, Jan. 2015. doi: 10.1109/JSSC.2014.2362842.

[25] SEMulator3D User's Guide, Version 8.0, Coventor, Inc., Rayleigh, NC, USA.

Chapter 4

A Comparison of Self-Heating Effects in Different Transistor Structures

4.1 Introduction

Self-heating effects (hereinafter referred to as “SHE”) are of concern for small-geometry metal-oxide-semiconductor (MOS) field-effect transistors (FETs) because they can degrade electrical performance [1] and reliability [2]. As transistor dimensions continue to be scaled down to achieve higher device densities, it is anticipated that the semiconductor industry will adopt nanosheet (also referred to as nano-ribbon [3] and multi-channel-bridge [4]) FET structures to provide for better electrostatic integrity and design versatility [5, 6] as compared with state-of-art FinFETs and are considered as a promising candidate for sub-5nm technology nodes [7]. Previous works focused on the intrinsic behaviors of self-heating in nanosheet FETs [8, 9, 10]. However, it is also important to understand how SHE may be worse for nanosheet FETs and whether they can still outperform FinFETs under the constraint of the same peak temperature.

In this chapter, we first compare n-channel and p-channel FinFETs (hereinafter referred to as “FF”) and nanosheet FETs (hereinafter referred to as “NSF”) in terms of self-heating and related device performance characteristics using Synopsys Sentaurus electro-thermal simulations [11]. Gate-all-around (GAA) FETs (hereinafter referred to as “GAAF”) with approximately square cross-section nanowire channel regions are also included for reference. This is because NSFs with small sheet widths are used in static memory (SRAM) cells. Then, based on the results of this preliminary study, the effect of various structural design parameters on SHE is studied in p-channel NSFs. Effective specific thermal resistance, R_{EFF}, is defined to gauge SHE and facilitate design optimization of NSFs. Design parameters varied include the nanosheet spacing (T_{SUS}), the raised source/drain region height (H_{SD}), the gate sidewall spacer thickness (L_{SP}), the source/drain length (L_{SD}), the sheet width, and the sheet pitch (SP), and source/drain Ge molefraction.

4.2 Simulation Methodology

Since the power of the generated heat is directly related to the current and voltage, all transistors under comparison should be designed so that they operate at the same supply voltage and have similar on-state current. Hence, to ensure a fair comparison between FF and NSF, it is critical to have a set of well-defined guidelines:

1. The NSF parameters are mostly adapted from [12].
2. The FF fin height should be the same as the NSF total stack height; we assume a fixed vertical etching capability in the same technology node.
3. The FF uses multiple (in this study, 4) fins to (approximately) match the effective channel width in NSF. Hence, the sheet pitch should be a multiple of the fin pitch. This is because an NSF is inherently a wide device, which occupies the same footprint as a multi-fin FF. In practice, the sheet width and the sheet pitch are lithographically defined and can be continuously scaled.
4. Fin Pitch – Fin Width = Sheet Pitch – Sheet Width. This is to ensure the same amount of metal, spacer materials, source/drain materials, etc. on the vertical sidewalls of fins or sheets so that they don't affect temperature calculations.
5. The normalized (per layout width) transistor off-state leakage current should be the same.
6. A control GAAF (with approximately square cross-section nanowires) is also included. The GAAF features 4 wires (as opposed to 3 sheets in NSF) and each wire is 7nm thick, so the total silicon stack is much higher. This design ensures comparable on-state current compared to FF and NSF.

Key transistor design parameter values for FF, NSF, and GAAF are summarized in Table 4.2-1.

	FF	**NSF**	**GAAF**
CPP (nm)	48		
FP/SP (nm)	16	64	16

L_G (nm)	12		
L_{SP} (nm)	6		
L_{SD} (nm)	12		
H (nm)	42	5	7
T_{SUS} (nm)	N/A	9	9
N_{FIN}	4	N/A	4
N_{WIRE}	N/A	3	4
Total Height (nm)	42	42	64
H_{SD} (nm)	20		
W (nm)	6	54	6
Total Width (nm)	360	354	416
x in SD $Si_{1-x}Ge_x$	0 (N-Channel) / 0.5 (P-Channel)		
I_{OFF} (nA/µm)	1		
V_{DD} (V)	0.75		
Channel Stress (GPa)	+0.4 (N-Channel) / -2 (P-Channel)		

Table 4.2-1. Key design parameters for transistors.

The following lists the definitions of the terms used in Table 4.2-1.

- CPP: Contacted poly pitch.
- FP: Fin pitch.
- SP: Sheet pitch. This is the spacing (in the **layout width direction**) between nanosheets in adjacent devices. This quantity should not be confused with the T_{SUS}, which is the **height** difference between adjacent nanosheets in the same device.
- L_G: Gate length.
- L_{SP}: Gate sidewall spacer length.
- L_{SD}: Source/Drain length.

- H: Silicon height. In FF, this is the fin height (= total height). In GAAF/NSF, this is the height of a **single** wire/nanosheet.
- T_{SUS}: Spacing between adjacent nanosheets/nanowires in the same transistor.
- N_{FIN}: Number of fins in a multi-fin FF and GAAF.
- N_{WIRE}: Number of wires per fin in a GAAF or number of nanosheets in a GAAF/NSF. For GAAF, $N_{WIRE} \times N_{FIN}$ = number of channels.
- Total Height: Height of the entire channel stack. In FF, this is the same as fin height. In GAAF/NSF, this is the height of all nanowires/nanosheets plus the spacing in between.
- H_{SD}: Height of source/drain over the top of the conductive channel.
- W: Silicon width. In FF, this is the fin width. In GAAF/NSF, this is the nanowire/nanosheet width.
- Total Width: total effective width. This corresponds to the total perimeter of all conductive channels in the transistor. In FF/GAAF, this number takes into account that a 4-fin transistor is used.
- I_{OFF}: Off-current specification. Measured at V_{GS}=V_{DS}=V_{DD}.
- V_{DD}: Supply voltage.

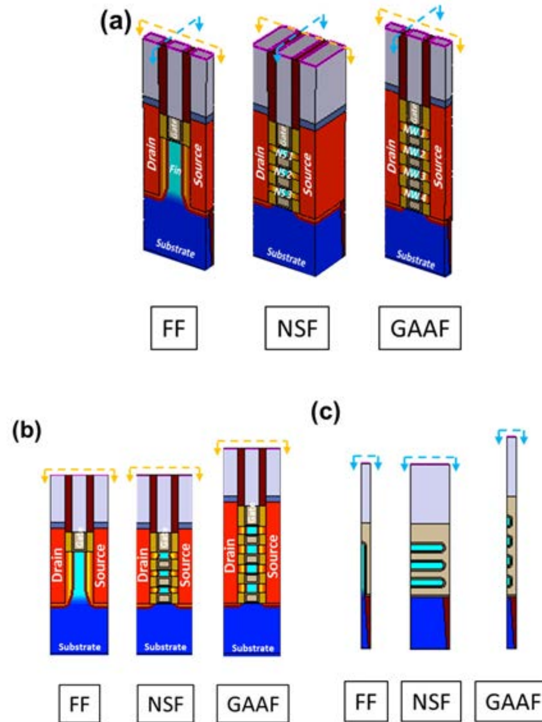


Figure 4.2-1. A comparison of the 3 (n-channel) transistor structures studied. (a) Isometric view, (b) Cross-gate cut view, (c) Cross-fin cut view.

The isometric and cross-sectional views of the simulated transistor structures are shown in Figure 4.2-1. To better show the conductive channel(s), the transistors are cut in half in the middle of the channel(s). Also note that the bottom portion of the substrate was omitted due to space limitation.

The transistor performance was simulated using the TCAD software package Sentaurus Device [11], using the drift-diffusion transport model with transport parameters calibrated to Monte-Carlo simulations, the inversion-accumulation layer model for carrier mobility [13] with thin-layer correction [14], the bandgap narrowing model, and the density gradient quantization model with parameters calibrated to empirical pseudopotential simulations [15]. The density gradient model is a computationally cheaper alternative to solving the full-blown self-consistent Poisson-Schrodinger equations [16]. The fin/nanosheet/nanowire vertical sidewall is assumed to be along {110} crystallographic planes, while the current flows in <110> direction. A +0.4GPa (tensile) and (-2GPa) compressive uniaxial stress along the conductive channel is assumed for n-channel and p-channel transistors, respectively. The thermodynamic model [17] is also turned on to assess SHE by solving lattice temperature equations alongside with the Poisson and continuity equations.

In order to save computational time, the multi-fin FF and GAAF are simulated by single fin FF and GAAF, respectively, with appropriate scaling factor. The validity of this approach is validated by comparing a real 4-fin FF and a 1-fin FF with the same normalized (to layout width) I_{OFF}. The comparison is listed in section 4.6.

4.3 Results and Discussion

4.3.1 N-Channel Transistors

Figure 4.3-1 and Table 4.3-1 summarize the electrical characteristics of the three n-channel transistors. L_{EFF} is the electrical gate length, defined as the distance between the two points in the channel that have a doping concentration of $2\times10^{19}cm^{-3}$. These three transistors have equal L_{EFF}; hence the difference of the electrostatics should mostly come from their structures [18]. As expected, N-FF has the worst SS because it is a double-gate structure, while the rest two are "gate-all-around." N-NSF has similar SS to that of N-GAAF; this implies that for n-channel, N-GAAF does not provide for much extra electrostatics benefit in facilitating further device scaling as compared with N-NSF. I_{MAX} (i.e., the

absolute value of the on-state current) is extracted at $V_{GS}=V_{DS}=V_{DD}$ and normalized to FP/SP. It can be seen that N-GAAF has the highest I_{MAX}, but achieving so at a much larger total conductive width.

Table 4.3-2 lists the average electron and current densities in each conductive channel. It can be seen in N-GAAF, the bottom nanowire (channel 4) has the lowest current density due to increased access resistance in the source/drain. The access resistance effect is not apparent in N-NSF as N-NSF has a much smaller total height.

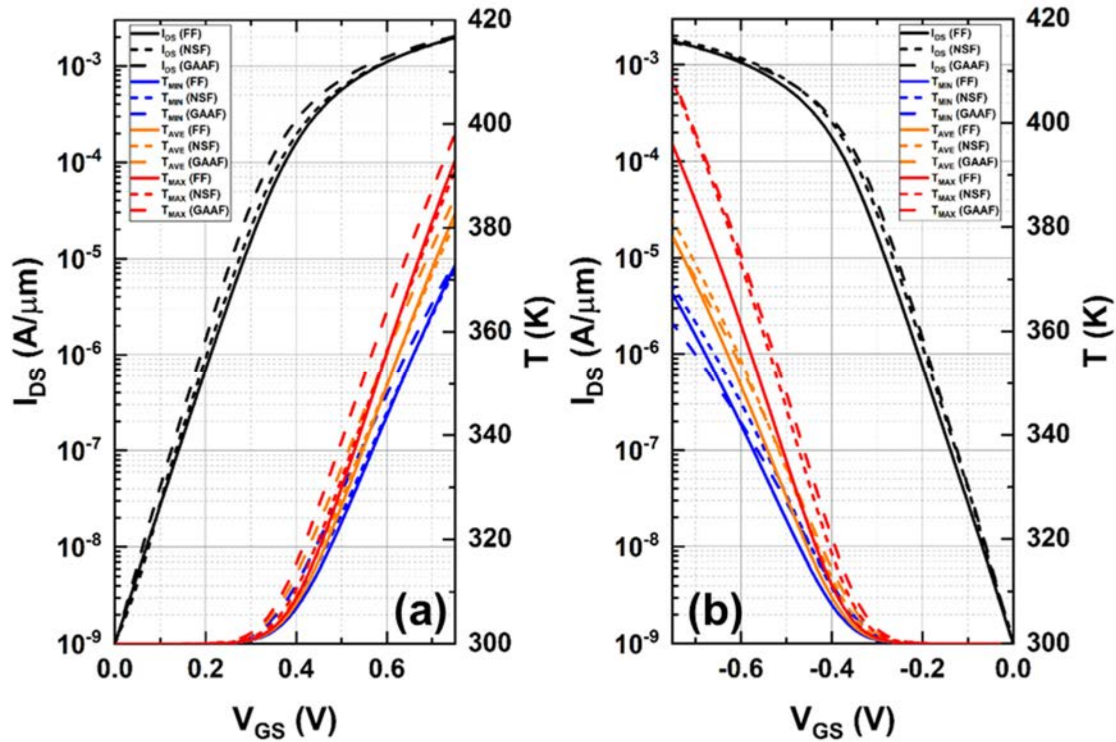


Figure 4.3-1. Simulated transistor current and temperatures as a function of gate voltages: (a) N-channel, (b) P-channel.

	L_{EFF} (nm)	Total Width (nm)	V_{TSAT} (V)	SS_{SAT} (mV/dec)	I_{MAX} (mA/μm)
N-FF	20	360	0.28	75	1.96
N-NSF	20	354	0.27	70	1.92
N-GAAF	20	416	0.25	69	2.02

Table 4.3-1. Electrical characteristics of n-channel transistors.

	e^- Density (cm^{-3})			**Current Density (A/cm^2)**		
	FF	**NSF**	**GAAF**	**FF**	**NSF**	**GAAF**
Channel 1	8.41×10^{18}	1.92×10^{19}	2.41×10^{19}	1.26×10^{7}	1.41×10^{7}	2.30×10^{7}
Channel 2	N/A	1.96×10^{19}	2.40×10^{19}	N/A	1.65×10^{7}	2.20×10^{7}
Channel 3	N/A	1.94×10^{19}	2.33×10^{19}	N/A	1.61×10^{7}	2.12×10^{7}
Channel 4	N/A	N/A	2.29×10^{19}	N/A	N/A	2.05×10^{7}

Table 4.3-2. Comparison of electron densities and current densities in the conductive channel(s) (n-channel).

	I_{MAX} (mA/μm)	T_{MIN} (K)	T_{AVE} (K)	T_{MAX} (K)
N-FF	1.96	373	383	393
N-NSF	1.92	372	381	391
N-GAAF	2.02	374	386	398

Table 4.3-3. Simulated n-channel transistor currents and temperatures.

In Table 4.3-3, T_{MIN}, T_{AVE}, and T_{MAX} represent the minimum, average, and maximum temperatures in the transistor channel regions, respectively. T_{MIN}, T_{AVE}, and T_{MAX} together provide a convenient means for comparing self-heating effects in the three transistor structures. It can be seen that for n-channel transistors, the three structures have similar temperatures. In addition, the temperatures in these three n-channel transistors are well correlated to the temperature; that is, a larger I_{MAX} leads to a higher temperature.

(a)	**FF**	**NSF**	**GAAF**
Channel 1	382	387	394
Channel 2	N/A	387	394
Channel 3	N/A	386	394
Channel 4	N/A	N/A	392

(b)	**FF**	**NSF**	**GAAF**
Channel 1	386	389	396
Channel 2	N/A	390	397
Channel 3	N/A	388	396
Channel 4	N/A	N/A	394

(c)	**FF**	**NSF**	**GAAF**
Channel 1	393	390	398
Channel 2	N/A	391	398
Channel 3	N/A	390	397
Channel 4	N/A	N/A	395

Table 4.3-4. Comparison of (a) T_{MIN}, (b) T_{AVE}, and (c) T_{MAX} in conductive channel regions (n-channel). Units in K.

Table 4.3-4 lists the minimum, average, and maximum temperatures in the conductive channel regions. Comparing with the numbers listed in Table 4.3-3, it can be concluded that T_{MAX} always occur in the conductive channels. And for N-NSF and N-GAAF, channels with larger on-state currents have larger T_{MAX}.

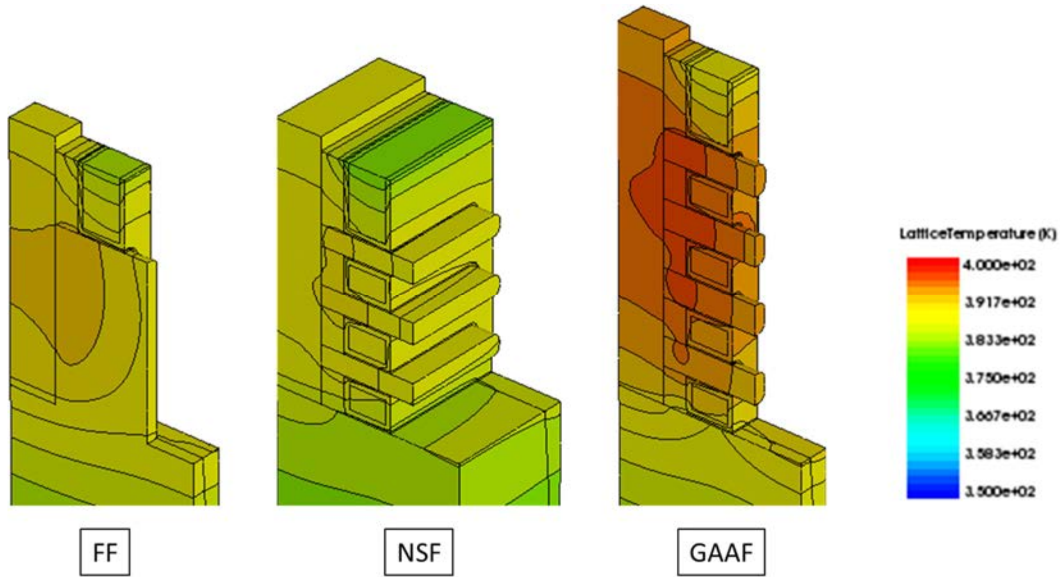


Figure 4.3-2. Temperature contours in n-channel transistors.

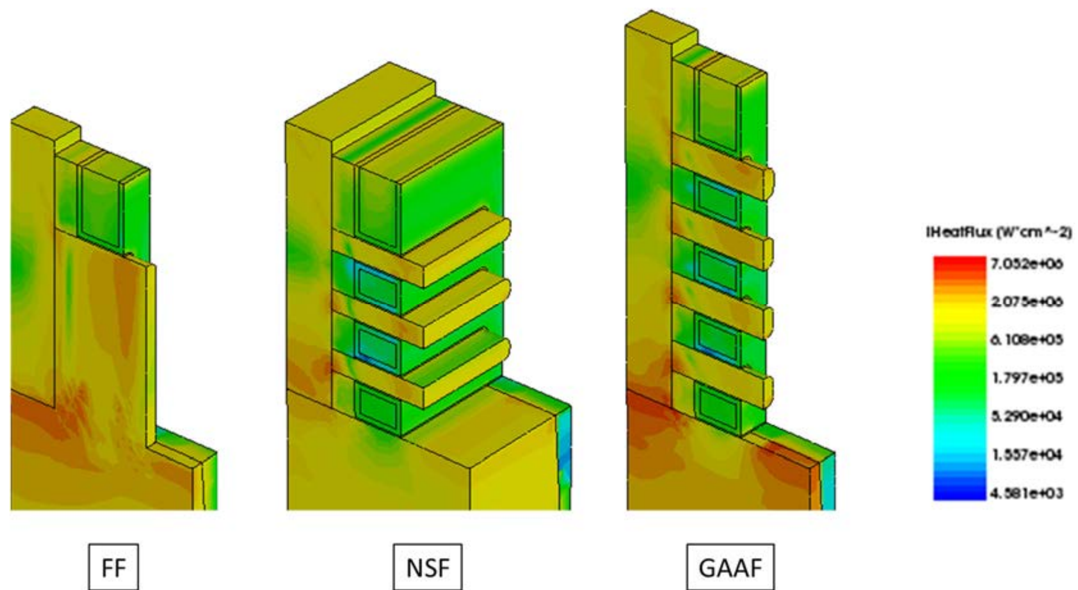


Figure 4.3-3. Heat flux in n-channel transistors.

Figure 4.3-2 and Figure 4.3-3 show the temperature distributions and heat flux in the transistors. For better clarity, the source region is omitted in these plots. From Figure 4.3-2, in N-FF, the temperature profile is smooth in the channel because only silicon is involved; this could potentially reduce the maximum temperature. In N-NSF, the highest temperature occurs near the boundary between drain and the channels. Compared with N-FF, the heat generated inside the nanosheets (channels) cannot easily propagate to the cooler

regions (e.g., substrate) easily since the materials in between (e.g., gate dielectrics, spacer materials) are poor thermal conductors. This situation is exacerbated in N-GAAF in which the current is more concentrated and the nanowires have smaller geometry. From Figure 4.3-3, we can see the heat is mostly dissipated towards the substrate from the source/drain-substrate interface. This is likely due to the fact in n-channel transistors, source, drain, and substrate are all silicon. It can also be seen that not much heat propagates through the poor thermal conductors (e.g., spacer materials and gate dielectrics).

4.3.2 P-Channel Transistors

Table 4.3-5 lists the electrical characteristics of the three p-channel transistors. Similar to the case of n-channel transistors, P-FF has the worst SS, while P-NSF and P-GAAF have similar SS. P-GAAF has the least on-state current even with the best SS and the largest total silicon width. P-NSF has larger on-state current than P-FF mostly due to better gate control.

	L_{EFF} (nm)	Total Width (nm)	V_{TSAT} (V)	SS_{SAT} (mV/dec)	I_{MAX} (mA/µm)
P-FF	20	360	-0.28	74	1.76
P-NSF	20	354	-0.26	68	1.85
P-GAAF	20	416	-0.25	68	1.70

Table 4.3-5. Electrical characteristics of n-channel transistors.

	h^+ Density (cm^{-3})			Current Density (A/cm^2)		
	FF	**NSF**	**GAAF**	**FF**	**NSF**	**GAAF**
Channel 1	6.14×10^{18}	1.23×10^{19}	1.51×10^{19}	1.13×10^{7}	1.52×10^{7}	1.88×10^{7}
Channel 2	N/A	1.24×10^{19}	1.53×10^{19}	N/A	1.45×10^{7}	1.87×10^{7}
Channel 3	N/A	1.22×10^{19}	1.48×10^{19}	N/A	1.42×10^{7}	1.80×10^{7}
Channel 4	N/A	N/A	1.44×10^{19}	N/A	N/A	1.75×10^{7}

Table 4.3-6. Comparison of electron densities and current densities in the conductive channels (p-channel).

	I_{MAX} (mA/µm)	T_{MIN} (K)	T_{AVE} (K)	T_{MAX} (K)
P-FF	1.76	367	378	396
P-NSF	1.85	369	381	408
P-GAAF	1.70	361	378	409

Table 4.3-7. Simulated p-channel transistor current and temperatures.

Table 4.3-6 records the hole and current densities in different conductive channels.

Table 4.3-7 lists the temperatures recorded in the transistors. It can be seen that P-GAAF has the worst self-heating since it has the highest T_{MAX} with the lowest I_{MAX}. This shows that for p-channel transistors, the transistor structure has a large impact on the self-heating. Compared with P-FF, P-NSF has a larger on-state current and also a larger T_{MAX}, so a further and more detailed analysis is required.

From Table 4.3-8, it can be seen that for all p-channel transistors, T_{MAX} is reached inside the conductive channel regions (as in the case of P-FF and P-NSF) or near the boundary between the drain and channels (as in the case of P-GAAF). We believe the latter one is an artifact caused by structure discretization. Compared with n-channel counterparts, p-channel transistors have significant hotter (higher T_{MIN}, T_{AVE}, and T_{MAX}) channels than their n-channel counterparts despite lower I_{MAX}.

As Figure 4.3-4 shows, there is a localized high temperature region centered at the boundary between drain and channels for all three p-channel transistors. For example, in N-FF (Figure 4.3-2), there is a continuous temperature contour from the drain to the fin (channel). As a comparison, in P-FF (Figure 4.3-4), there is a high temperature island centered at the drain and fin boundary. This is due to the fact that in p-channel transistors, a 50% Germanium mole-fraction SiGe is used to boost the stress in the channel. $Si_{0.5}Ge_{0.5}$ has much lower thermal conductivity to that of silicon [19]; hence the heat generated in the fin (channel) cannot be well dissipated to the substrate from the drain region.

(a)	FF	NSF	GAAF
Channel 1	379	400	398
Channel 2	N/A	402	403
Channel 3	N/A	394	400
Channel 4	N/A	N/A	389

(b)	FF	NSF	GAAF
Channel 1	384	404	401
Channel 2	N/A	406	406
Channel 3	N/A	398	403
Channel 4	N/A	N/A	394

(c)	FF	NSF	GAAF
Channel 1	396	407	404
Channel 2	N/A	408	408
Channel 3	N/A	400	405
Channel 4	N/A	N/A	396

Table 4.3-8. Comparison of (a) T_{MIN}, (b) T_{AVE}, and (c) T_{MAX} in conductive channel regions (p-channel). Units in K.

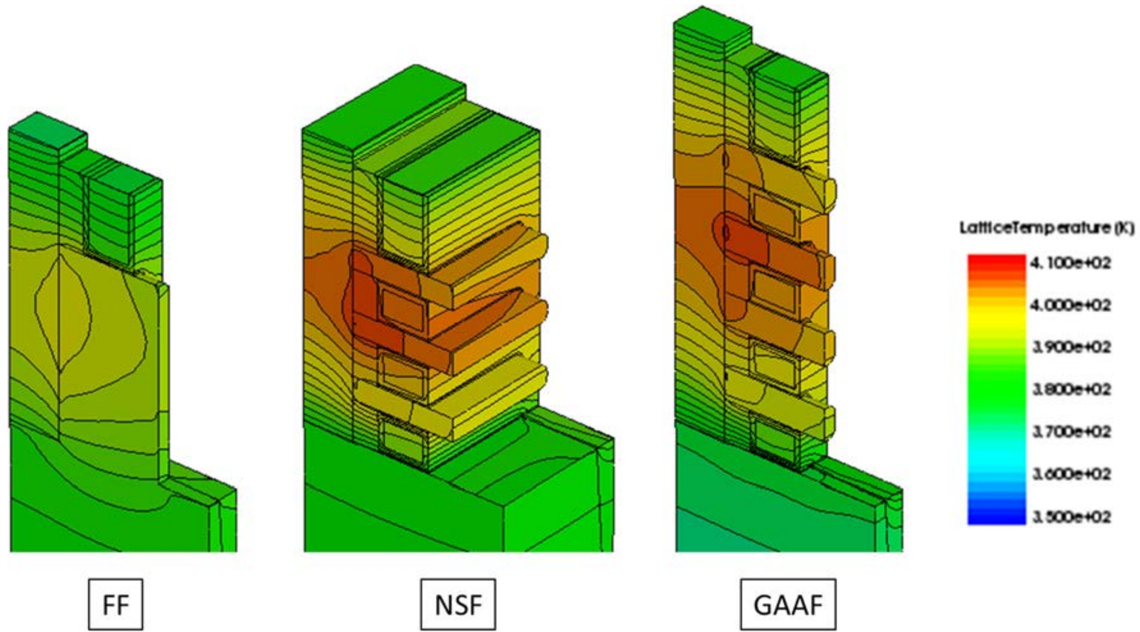


Figure 4.3-4. Temperature contours in p-channel transistors.

Figure 4.3-5 shows the heat flux inside the p-channel transistors. It can be seen that in P-FF, the heat flows mostly from the silicon fin to the substrate; there is much less heat flow from the drain to the substrate. In P-NSF and P-GAAF, the heat is dissipated from bottom nanosheet/nanowire the spacer materials down to the substrate. This is likely due to the fact that $Si_{0.5}Ge_{0.5}$ has lower thermal conductivity ($\varkappa = 0.088$W/cm·K) than the low-k spacer material ($\varkappa = 0.18$W/cm·K) [19, 20]. As a reference, Si has a much larger thermal conductivity ($\varkappa = 1.6$W/cm·K) [11]. It should be noted that in P-NSF and P-GAAF, the top and middle wires usually conduct more current and generate more heat. The heat generated in the upper nanosheets and nanowires is mainly dissipated through the spacers to the lower nanosheets and nanowires and ultimately to the substrate. In addition, for the top nanosheet and nanowire, it is also possible to dissipate the heat through the top spacer and the upper gate stack. This observation implies that adding more nanosheets and nanowires to p-channel NSFs and GAAFs is not optimal as it worsens the thermal heating in the middle nanosheets and nanowires.

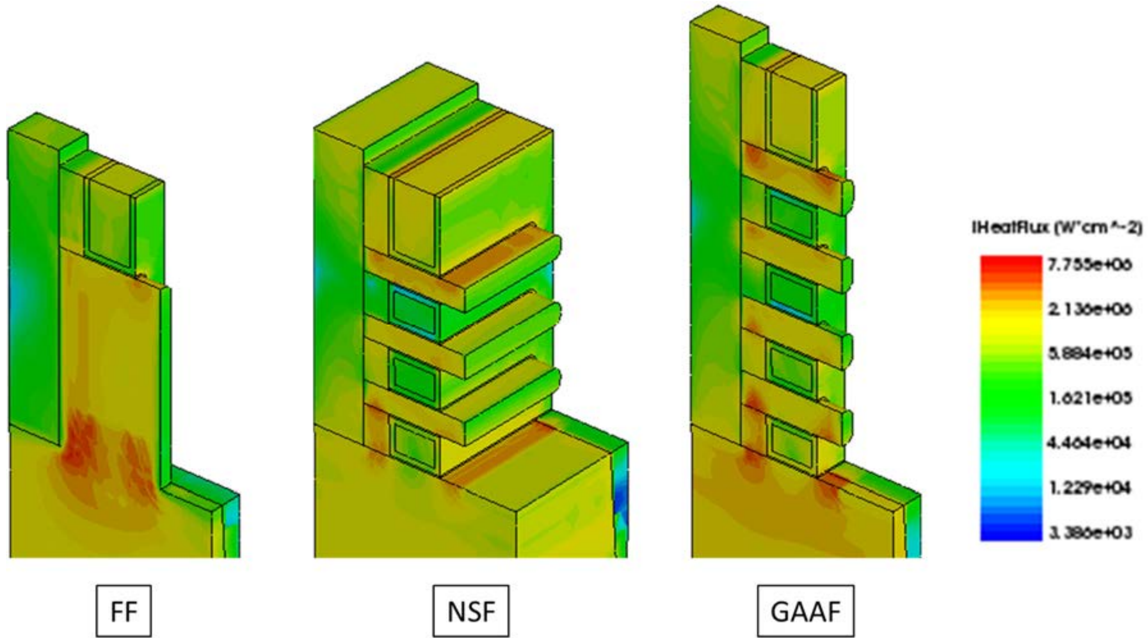


Figure 4.3-5. Heat flux in p-channel transistors.

4.3.3 Iso-T_{MAX} Performance Comparison

Since the hottest spot in a transistor is correlated to transistor reliability, in this section, we gradually lower the supply voltage (V_{DD}) for P-NSF and P-GAAF to match the T_{MAX} in P-FF and compare its on-state current to that of P-FF. N-NSF and N-GAAF share the same V_{DD} as their p-channel counterparts.

	$\|V_{DD}\|$ (V)	**I_{MAX} (mA/μm)**	**T_{MAX} (K)**
N-FF	0.75	1.96 (100%)	393
P-FF	0.75	1.76 (100%)	396
N-NSF	0.72	1.76 (90%)	381
P-NSF	0.72	1.72 (98%)	397
N-GAAF	0.71	1.83 (93%)	385
P-GAAF	0.71	1.57 (89%)	396

Table 4.3-9. V_{DD} scaling of NSFs and GAAFs to match T_{MAX} as in FFs. The percentages in parentheses are the current ratios to the corresponding FFs.

It can be seen from Table 4.3-9 that in order to match the maximum temperature, T_{MAX}, the V_{DD} of NSF needs to be lowered to 0.72V to bring T_{MAX} of P-NSF to 396K. This results in comparable current between P-FF and P-NSF. However, lowering V_{DD} reduces N-NSF on-state current to only 90% of that in an N-FF. For GAAFs, the V_{DD} is lowered further to 0.71V; this results in 11% reduction and 7% reduction in on-state current compared to P-FF and N-FF, respectively.

4.3.4 Summary

From the above discussions, we conclude that:

1. N-channel transistors have less self-heating than their p-channel counterparts.
2. In n-channel transistors, the maximum temperature (T_{MAX}) is mostly correlated to the current in the conductive channel(s); a larger current implies a higher T_{MAX}. The transistor structural difference does not play a significant role in affecting the temperatures.
3. Since a silicon source/drain is used in n-channel transistors, the heat generated in the channels can dissipate through the source/drain to the substrate. In N-FF, the heat can also flow from the silicon fin to the silicon substrate, resulting in a smooth temperature profile across the whole fin (channel). This helps reduce the T_{MAX} in N-FF.
4. P-channel transistors feature a $Si_{0.5}Ge_{0.5}$ source/drain to apply uniaxial compressive stress to the channel. However, since $Si_{0.5}Ge_{0.5}$ has $>$ 10X less thermal conductivity than that of silicon, the heat generated in the conductive channels cannot be well dissipated through the source/drain. In P-FF, the heat can still flow from the fin to the substrate. In P-NSF and P-GAAF, since the nanosheets and nanowires are suspended from the substrate, the bottom nanosheet and nanowire can only dissipate the heat through the bottom spacer to the silicon substrate. And the top nanosheet and nanowire can dissipate the heat through the spacer on top. The nanosheets and nanowires in the middle have the poorest thermal path to either the top or to the bottom. Hence, for similar on-state current, GAAFs are the hottest because a higher aspect ratio structure (e.g., more nanowires) for performance reasons. This implies adding more nanosheets

or nanowires is not preferable as it worsens the thermal situation in the middle wires.

5. Matching T_{MAX} requires lowering V_{DD} for P-NSF and P-GAAF, which degrades the performance (i.e., the on-state current of both p-channel and n-channel are worse).

4.4 Optimization of P-Channel NSF

In section 4.3, it is found that p-channel transistors have worse self-heating than their n-channel counterparts due to much lower thermal conductivity in $Si_{0.5}Ge_{0.5}$ (source/drain) than that of silicon. Moreover, both NSF and GAAF show worse SHE than FF since the middle nanosheets and nanowires have poor thermal paths to the substrate or to the top. As a result, to achieve the same T_{MAX} and reliability, the NSF and GAAF must lower their supply voltages. As shown in Table 4.3-9, the NSF and GAAF conduct less current than FF when they have the same T_{MAX}.

In this section, p-channel NSF is optimized under the constraint of self-heating by adjusting various transistor structural and material parameters and studying their effects on self-heating. Then the optimized P-NSF is presented to benchmark against the P-FF.

4.4.1 Optimization Methodology

It is understood that T_{MAX} is a more representative number to show the degree of self-heating as it correlates with the transistor reliability. But as seen from section 4.3, it is impacted by the conduction current (I_{MAX}) and the supply voltage (V_{DD}). It is also expected that the transistor maximum current (I_{MAX}) will change as various structural and material parameters are varied. In addition, T_{MAX} becomes smaller with a larger layout area and/or less generated power. To facilitate NSF design optimization, an effective specific thermal resistance, R_{EFF}, is proposed as the target of the optimization:

$$R_{EFF} = \frac{T_{MAX} - 300K}{P_A} \text{ (Equation 4.4-1)}$$

where T_{MAX} is the maximum temperature measured in the unit of K, P_A is the power generated per unit layout area due to Joule heating. And P_A can be written as,

$$\text{FF and GAAF: } P_A = \frac{V_{DD} I_{MAX}}{A} = \frac{V_{DD} I_{MAX}}{CPP \cdot FP}$$

$$\text{NSF: } P_A = \frac{V_{DD} I_{MAX}}{CPP \cdot SP} \quad \text{(Equation 4.4-2)}$$

A lower R_{EFF} is considered better (i.e., less self-heating).

As a sanity check, the results from section 4.3 are used to calculate R_{EFF} (Table 4.4-1). It can be seen that R_{EFF} is a good indicator in comparing the self-heating in different transistor structures. R_{EFF} is similar across all 3 n-channel transistors, re-affirming that the structural difference does not play a big role in self-heating. P-channel transistors have much larger R_{EFF} than their n-channel counterparts. And as expected from previous analysis, P-GAAF has the largest R_{EFF}, while P-FF has the smallest R_{EFF}.

	I_{MAX} (mA/μm)	**T_{MAX} (K)**	**R_{EFF} (K·cm²/W)**
N-FF	1.96	393	3.0×10^{-5}
N-NSF	1.92	391	3.0×10^{-5}
N-GAAF	2.02	398	3.1×10^{-5}
P-FF	1.76	396	3.5×10^{-5}
P-NSF	1.85	408	3.7×10^{-5}
P-GAAF	1.70	409	4.1×10^{-5}

Table 4.4-1 Calculated R_{EFF} for different transistor structures in section 4.3.

In addition to R_{EFF}, a pure thermal simulation approach is also proposed to achieve much faster turn-around time and qualitative understanding in relevant cases. In this approach, only the heat diffusion equation is solved. Based on the study in section 4.3, maximum temperature spots are located in the vicinities of the channels near the drain regions. Therefore, the heat source is placed at the drain side of the channel with total thermal power generated based on the P_A calculated based on results from section 4.3. Figure 4.4-1 shows the placement of the thermal source in NSF.

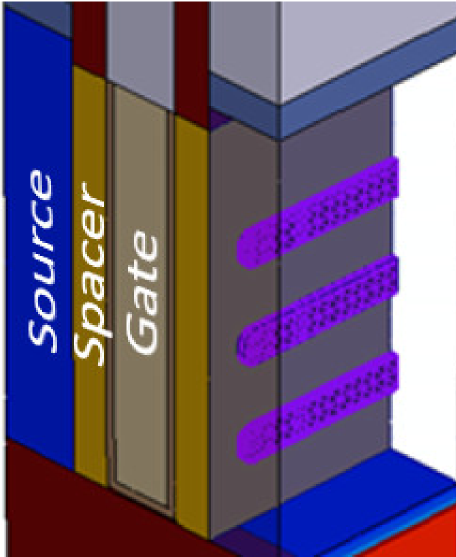


Figure 4.4-1. Thermal contact at drain-side.

It should be noted that due to the limitations of the constant-power assumed in these pure thermal simulations, they may not be appropriate for parameters that can cause large change in current and heat.

4.4.2 Effect of Nanosheet Spacing (T_{SUS})

Nanosheet spacing (nominal: 9nm) is varied from 5nm to 50nm. Note that nanosheet spacing is defined to be the distance between the top surface of the lower sheet to the bottom surface of the higher sheet; hence it includes the thicknesses of gate stack in between (Figure 4.4-2). For example, a 5nm T_{SUS} consists of 2×0.4nm=0.8nm interfacial oxide layer, 2×1.28nm=2.56nm high-k dielectric, leaving only ~1.6nm for the workfunction gate metal layers and the tungsten filling, which would be challenging in practice.

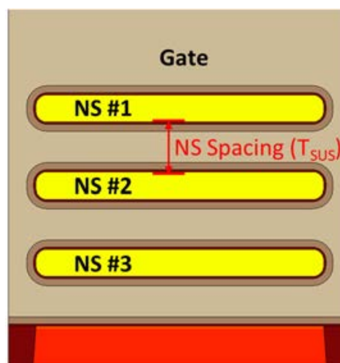


Figure 4.4-2. Definition of Nanosheet Spacing.

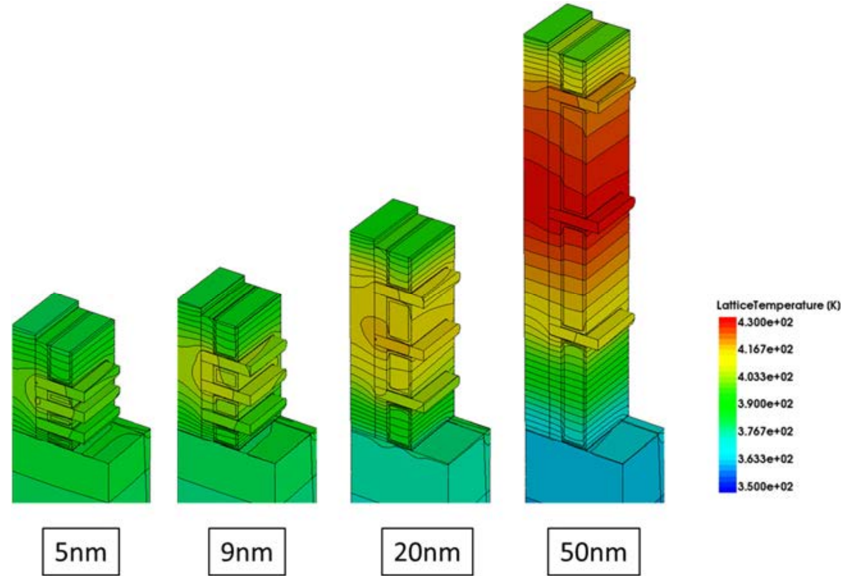


Figure 4.4-3. Temperature contours for various T_{SUS}.

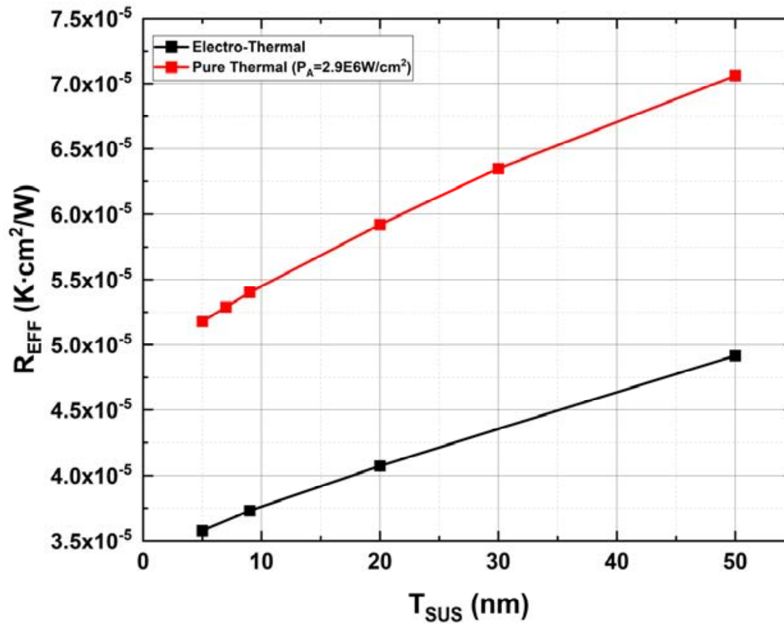


Figure 4.4-4. R_{EFF} calculated from pure thermal and electro-thermal simulation as a function of T_{SUS}.

Figure 4.4-3 shows the temperature contours for various T_{SUS} cases. It can be seen that the middle nanosheet has the highest T_{MAX} in all cases. In addition, the thermal situation worsens when T_{SUS} increases. This is because, at large spacing, the heat has to travel longer path in the S/D region to reach the thermal sinks

and the heat from the middle sheet cannot dissipate well through the substrate. Note that the substrate is getting cooler in larger T_{SUS} cases. It is also found that the top nanosheet has a higher T_{MAX} than the bottom nanosheet when the nanosheet spacing is large. This indicates that heat dissipates through the substrate easier than the top source/drain/gate contacts at the top. Figure 4.4-4 also shows that pure thermal simulation and electro-thermal simulation gives the same trend.

4.4.3 Effect of Raised Source/Drain Height (H_{SD})

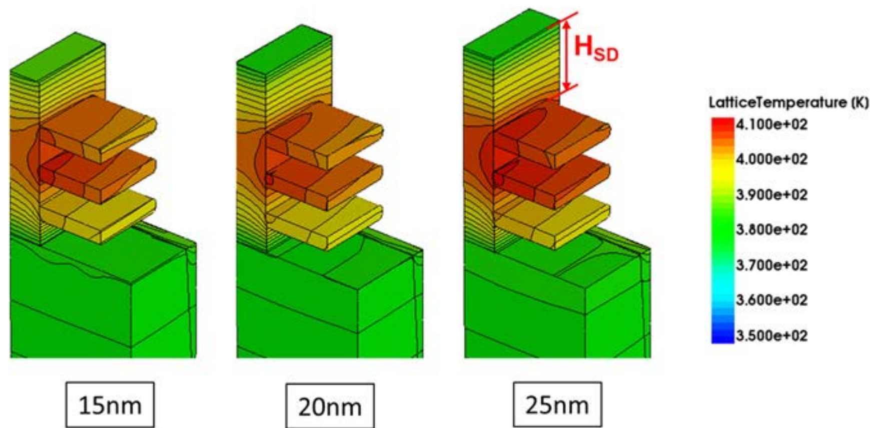


Figure 4.4-5. Temperature contours for various H_{SD}.

In the nominal case from section 4.3, all the structures have a raised source and drain design (i.e., the source and drain top is above the top surface of the conductive channel). In this study, the source/drain height is varied from 15nm to 25nm to study its effect on SHE. The nominal value is 20nm. As shown in Figure 4.4-5 and Figure 4.4-6, the source/drain height variation has a negligible impact on the thermal resistance. This further confirms the previous discussion that heat dissipates mostly downwards. It is also found that pure thermal simulation gives a similar trend as electro-thermal simulation.

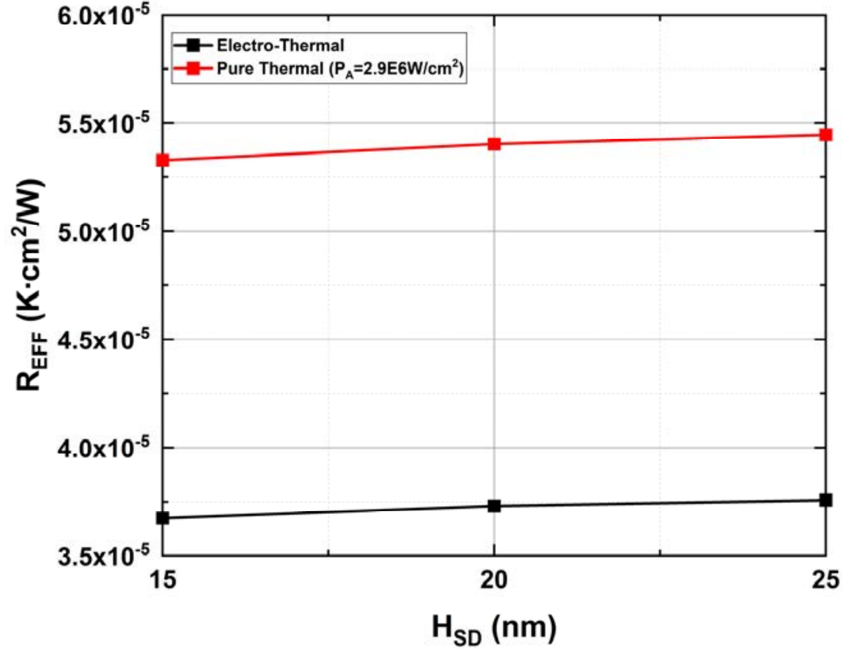


Figure 4.4-6. R_{EFF} calculated from pure thermal and electro-thermal simulation as a function of H_{SD}.

4.4.4 Effect of Sheet Pitch and Sheet Width

In this section, we consider two scenarios: (1) the sheet pitch is varied while the sheet width is kept constant, and (2) the difference between the sheet pitch and the sheet width is kept constant while the sheet width is varied.

The first scenario explores the optimal difference between the sheet pitch and the sheet width (i.e., the optimal width of the sidewall spacers, STI, etc.). A larger sheet pitch with fixed sheet width implies a wider STI in general. The pure simulation is appropriate in this scenario as the total current should not change much due to the fixed sheet width.

Figure 4.4-7 shows the temperature distributions of NSF with 3 different sheet pitches. It can be seen that NSF has lower T_{MAX} when the sheet pitch increases. However, Figure 4.4-8 shows R_{EFF} increases when sheet pitch increases. This is because when the sheet pitch increases, the additional area is covered by oxide in STI, which has a much lower thermal conductivity than silicon. R_{EFF} is worse due to the waste of area.

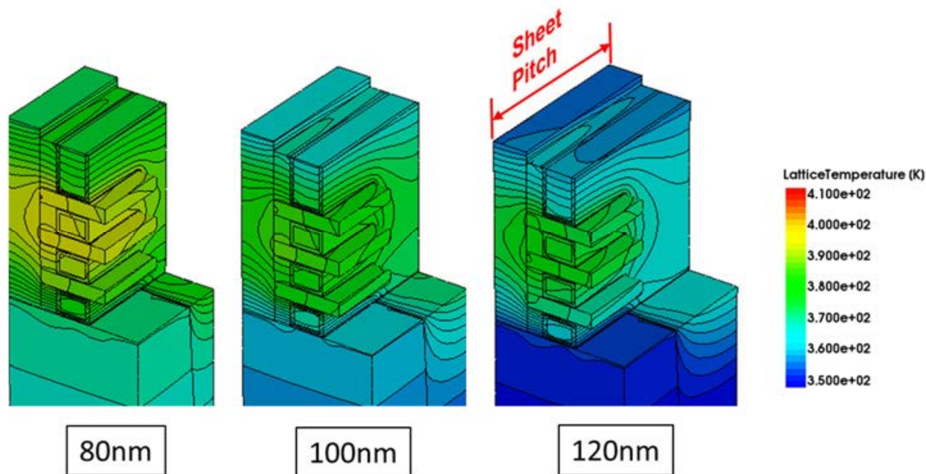


Figure 4.4-7. Temperature contours for various SP.

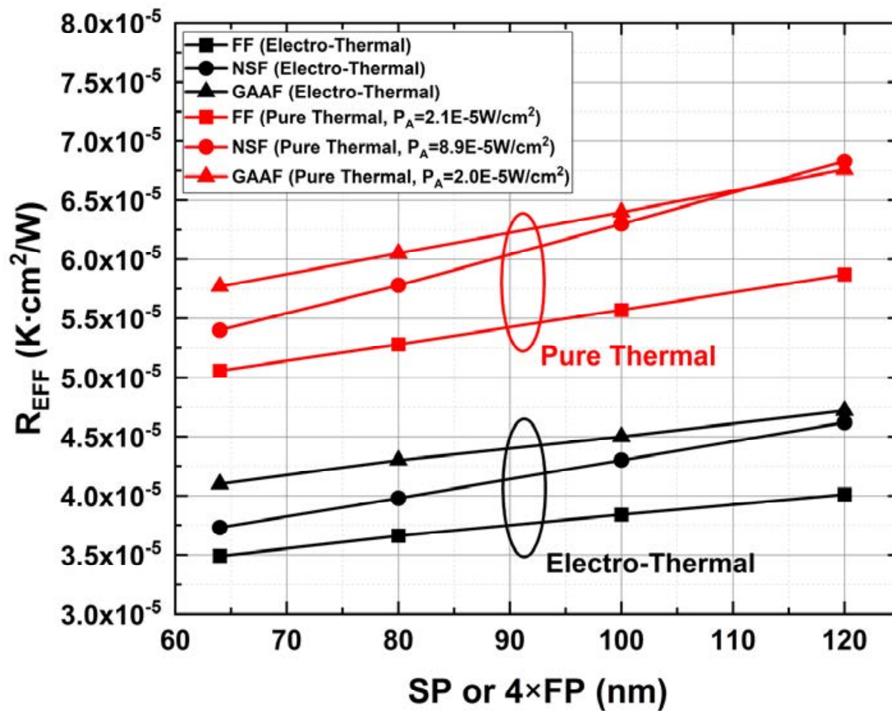


Figure 4.4-8. R_{EFF} calculated from pure thermal and electro-thermal simulations as a function of pitch. In FF and GAAF, pitch=4×FP; In NSF, pitch=SP. The specific power (P_A) for each transistor structure used in the pure thermal simulation is calculated based on I_{MAX} and V_{DD} in section 4.3.

In addition, as a reference, the fin pitch of FF and GAAF is also varied in the same manner and their effects are also studied. As mentioned in section 4.2,

to ensure a fair comparison, the sheet pitch should always be a multiple of the fin pitch. In this study, we fix this ratio to 4. That is, for a given sheet pitch, 4-fin FF or 4-fin GAAF can be placed evenly. It is found that FF still has the best R_{EFF} while GAAF is the worst (this was found in section 4.3 already). Moreover, NSF has the worst slope as the pitch increases. This is because NSF has only one STI opening for the entire structure; while in FF and GAAF, there are 4 STI openings. Therefore, although it has the same silicon area as FF and GAAF, NSF enjoys much less heat spreading effect. The pure thermal simulation gives similar results as electro-thermal simulation as shown in Figure 4.4-8.

One advantage of NSF is that its sheet width can be defined by the lithography, which can be (theoretically) scaled continuously. In the second scenario, the difference between the sheet pitch and the sheet width is fixed while the sheet width is varied. By doing so, we can examine how sensitive SHE is to the sheet width in the same NSF fabrication process. Since the current varies much, the pure thermal simulation is not appropriate here.

Sheet Width (nm)	**I_{MAX} (mA/µm)**	**T_{MAX} (K)**	**R_{EFF} (K·cm²/W)**
32	1.54	394	3.9×10^{-5}
48	1.68	400	3.8×10^{-5}
64	1.85	408	3.7×10^{-5}

Table 4.4-2 R_{EFF} for different NSF sheet width (sheet pitch – sheet width constant).

It can be seen from Table 4.4-2, that NSF with a larger sheet width is better in both the current density and the SHE. This is because for the same difference between the sheet pitch and the sheet width, a larger sheet width implies a higher substrate to STI area ratio, and hence a better heat transfer from the nanosheets to the substrate.

4.4.5 Effect of Spacer Length (L_{SP})

In this section, L_{SP} is varied while L_G and L_{SD} are fixed.

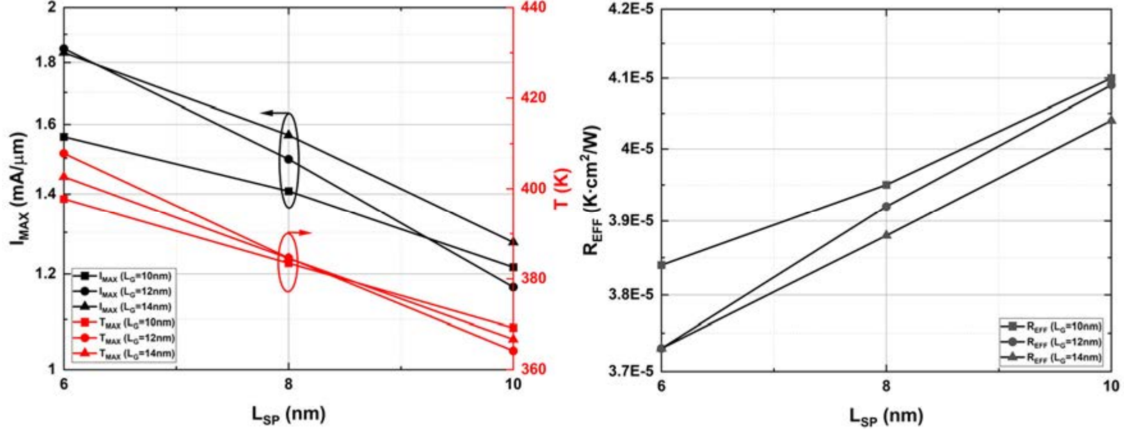


Figure 4.4-9. Left: I_{MAX} and T_{MAX} as a function of L_{SP}. Right: R_{EFF} as a function of L_{SP}.

As shown in the left of Figure 4.4-9, for all gate lengths, reducing L_{SP} increases the on-state current and the maximum temperature. From the right of Figure 4.4-9, we can see that a smaller L_{SP} leads to a smaller R_{EFF}. In addition, a smaller L_G also has a larger R_{EFF} when having the same L_{SP}; this is due to the degradation of electrostatics at smaller L_G. Note that the pure simulation is not appropriate here as the current varies by a large margin.

4.4.6 Effect of Source/Drain Length (L_{SD})

In this section, the effects of contacted poly pitch of NSF on SHE are explored by varying L_{SD}. Note that CPP and FP/SP are critical numbers that characterize the process as it directly relates to the density of the transistors. However, it is still worthwhile to investigate whether by tweaking L_{SD} within a small window, the self-heating can be improved.

L_{SD}/CPP (nm)	I_{MAX} (mA/µm)	T_{MAX} (K)	R_{EFF} (K·cm²/W)
10/44	1.63	403	3.7×10^{-5}
12/48 (Reference)	1.85	408	3.7×10^{-5}
14/52	1.88	404	3.8×10^{-5}

Table 4.4-3 R_{EFF} for different source/drain length (L_{SD}).

From Table 4.4-3, we can see that a longer source and drain can increase the on-state current. This is because a longer source and drain has a larger surface

area which reduces the contact-to-source/drain resistance. Also note that the pure thermal simulation is not appropriate here as it does not account for the change in on-state current.

There is also a competing factor: as shown in Figure 4.4-10, a longer source and drain can help reduce the maximum temperature near the boundary of drain and the second nanosheet. This is because the heat generated at the boundary has larger room to spread in the length direction, effectively reducing the maximum temperature.

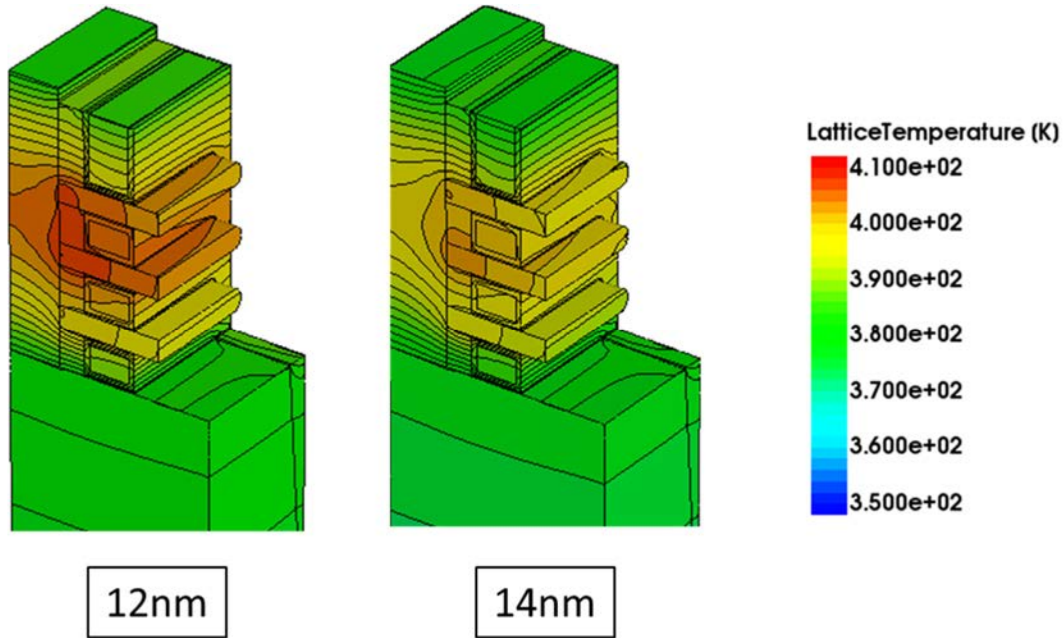


Figure 4.4-10. Temperature contours with various L_{SD}. Only 12nm (default) and 14nm L_{SD} case are shown as they have similar current.

However, since R_{EFF} takes the layout area (CPP×SP) into account, at some point, the minor change in source/drain resistance and the improvement in heat crowding do not worth the increase of CPP. As shown in Table 4.4-3, R_{EFF} does not vary much when L_{SD} increases from 10nm to 14nm (+40%), or CPP increases from 44nm to 52nm (+18%). Therefore, tweaking L_{SD} within a small range does not help reduce self-heating in NSF.

4.4.7 Effect of Gate Sidewall Spacer Length (L_G) and Spacer Length (L_{SP})

In this section, CPP and L_{SD} are fixed at 48nm and 12nm, respectively, while L_G and L_{SP} are varied. More specifically, we look at 3 combinations of (L_G, L_{SP}): (12nm, 6nm), (10nm, 7nm), and (14nm, 5nm). The doping gradient are fixed and hence these three designs have similar L_{EFF}.

(L_G, L_{SP}) (nm)	**I_{MAX} (mA/μm)**	**T_{MAX} (K)**	**R_{EFF} (K·cm²/W)**
(10, 7)	1.47	390	3.9×10^{-5}
(12, 6) (Reference)	1.85	408	3.7×10^{-5}
(14, 5)	1.96	413	3.7×10^{-5}

Table 4.4-4 R_{EFF} for different L_G and L_{SP} combinations.

(10nm, 7nm) case has the largest R_{EFF} due to the lowest on-state current (Table 4.4-4). This is due to a worse electrostatics resulted from a shorter gate length (Figure 4.4-11). The (14nm, 5nm) and the reference (12nm, 6nm) has similar R_{EFF}.

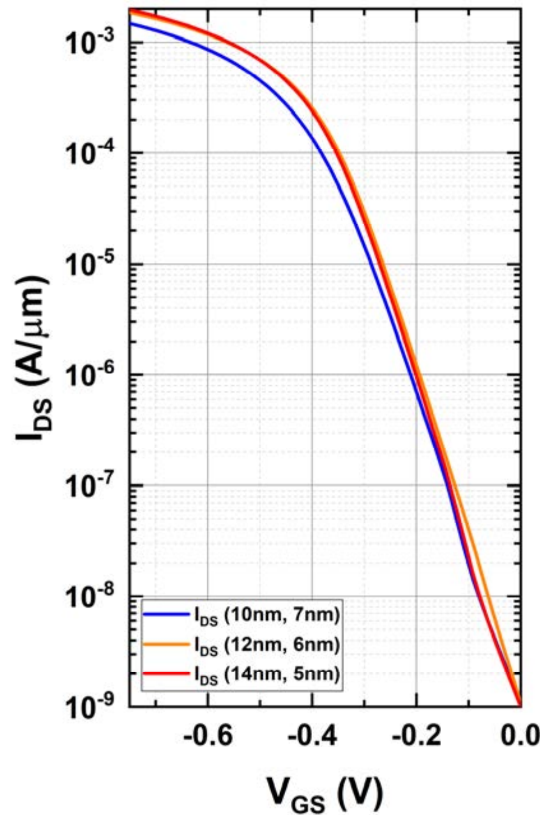


Figure 4.4-11. Simulated transfer characteristics for different (L_G, L_{SP}) combinations.

4.4.8 Effect of Gate Sidewall Spacer Length (L_{SP}) and S/D Length (L_{SD})

In this study, CPP and L_G are fixed at 48nm and 12nm, respectively, while L_{SP} and L_{SD} are varied. In particular, three combinations of (L_{SP}, L_{SD}) are investigated: (6nm, 12nm), (8nm, 10nm), and (10nm, 8nm). To maintain the same leakage current, the source/drain doping gradients are adjusted so they have the same L_{EFF} (extracted at $2\times10^{19}cm^{-3}$ doping concentration) as shown in Figure 4.4-12.

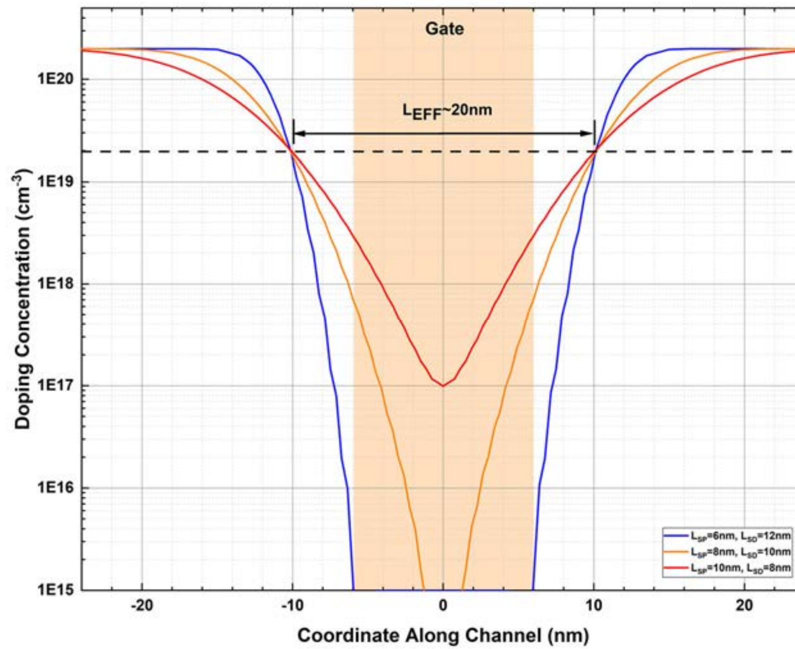


Figure 4.4-12. Doping concentration profiles in various (L_{SP}, L_{SD}) combinations. The effective channel length (extracted at $2\times10^{19}cm^{-3}$ doping concentration) is kept the same by tuning the doping gradient from the source/drain region.

(L_{SP}, L_{SD}) (nm)	**I_{MAX} (mA/μm)**	**T_{MAX} (K)**	**R_{EFF} (K·cm²/W)**
(6, 12) (Reference)	1.85	408	3.7×10^{-5}
(8, 10)	1.78	403	3.7×10^{-5}
(10, 8)	1.64	396	3.7×10^{-5}

Table 4.4-5 R_{EFF} for different L_{SP} and L_{SD} combinations.

Table 4.4-5 lists the computed R_{EFF} for these three cases. The change in R_{EFF} is minimal. Note that the degradation in I_{MAX} in the (10, 8) case compared to (6, 12) case is due to non-negligible doping in the channel region, causing mobility degradation and hence lower on-state drive current. Also note that in this study I_{MAX} varies, and hence pure thermal simulations are not appropriate as they assume a constant-power source-side thermal contact.

4.4.9 Effect of Source/Drain Germanium MoleFraction

In this study, the effect of source/drain $Si_{1-x}Ge_x$ material is studied. While the channel is fixed to pure silicon, the source/drain $Si_{1-x}Ge_x$ is varied from x = 0 (pure Silicon) to x = 1 (pure Germanium). The nominal x is 0.5. The compressive stress induced in the Si channel is also varied linearly from 0 to -4GPa to account for the change of mole fraction in the source/drain. The gate workfunction is tuned so that each transistor has an I_{OFF}=1nA/µm.

Ge Molefraction	**I_{MAX} (mA/µm)**	**T_{MAX} (K)**	**R_{EFF} (K·cm²/W)**
0 (Pure Si)	1.23	364	3.2×10^{-5}
0.2	1.61	396	3.8×10^{-5}
0.4	1.78	405	3.8×10^{-5}
0.5 (Reference)	1.85	408	3.7×10^{-5}
0.6	1.85	408	3.7×10^{-5}
0.8	1.61	395	3.8×10^{-5}
1.0 (Pure Ge)	1.13	363	3.6×10^{-5}

Table 4.4-6. Simulated transistor current and temperatures for different source/drain Ge molefraction.

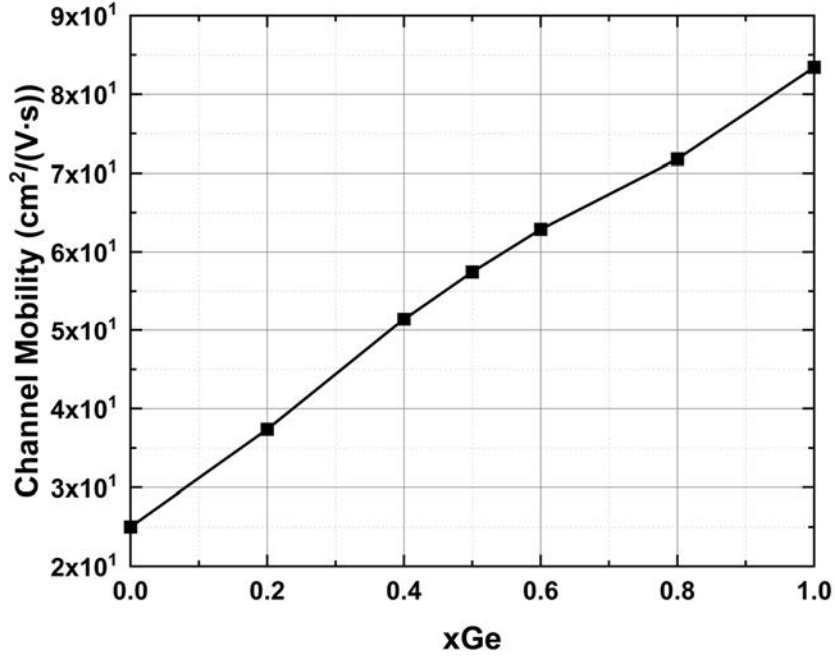


Figure 4.4-13. Channel average mobility as a function of source/drain Ge molefraction (V_{GS}=V_{DS}=V_{DD}).

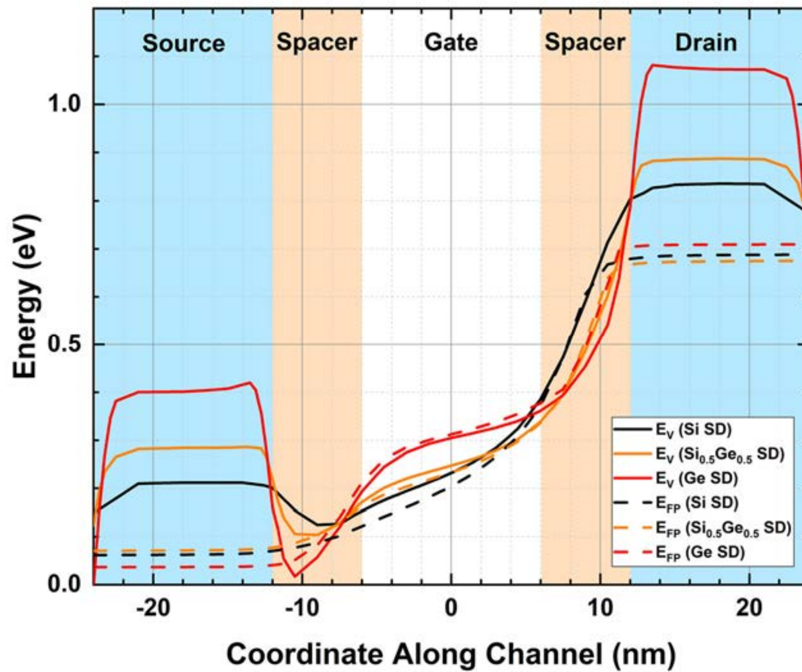


Figure 4.4-14. Simulated valance band and hole quasi-fermi level profile along the channel (V_{GS}=V_{DS}=V_{DD}). The source-side barrier is at around x=-10nm.

From Table 4.4-6, when the source/drain Ge molefraction (x) is around 0.5, the transistor achieves the highest I_{MAX}. There are two factors. Firstly, as shown in Figure 4.4-13, as x increases from 0 to 1 (from pure Si to pure Ge), the transistor enjoys higher mobility (linear-regime) and higher velocity (saturation-regime) in the channel due to the higher compressive stress. Secondly, as shown in Figure 4.4-14, using $Si_{1-x}Ge_x$ source/drain with larger x can cause increased source-side barrier due to the $Si_{1-x}Ge_x$-Si heterojunction. This reduces the number of holes or electrons diffusing into the channel in the on-state. Since the current is proportional to the inversion charges, this eventually leads to a reduction of the on-state current.

In addition, other non-ideal factors might also play a role. For example, a higher x $Si_{1-x}Ge_x$ will have a decreased density-of-state.

It should also be clear that from Table 4.4-6, pure Si has the best R_{EFF} and pure Ge the second. And the rest using $Si_{1-x}Ge_x$ has larger R_{EFF} than these two. This is well expected: pure Si has the largest thermal conductivity and pure Ge has the second largest thermal conductivity. Most $Si_{1-x}Ge_x$ alloy has about 10X less thermal conductivity than that of pure Si, and 5X less than that of pure Ge (Figure 4.4-15). As seen from Figure 4.4-16 (note the different scales for each plot), most of the heat is dissipated into drain-to-substrate boundary in the pure Si source/drain case. In the $Si_{0.5}Ge_{0.5}$ source/drain case, the heat is mostly dissipated through the bottom spacers to the substrate. And in the pure Ge source/drain case, a portion of heat is dissipated through the drain-to-substrate boundary, while the rest flows through the bottom spacers.

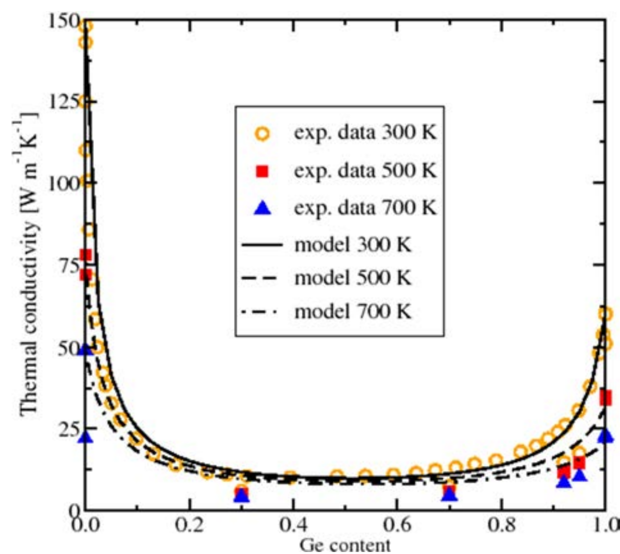


Figure 4.4-15. Thermal conductivity of Si, SiGe, and Ge [21].

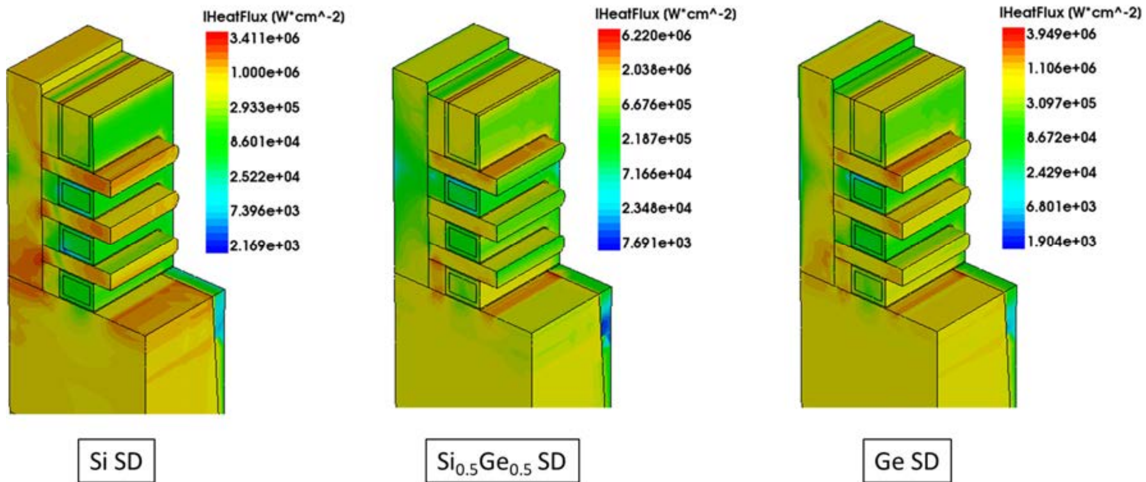


Figure 4.4-16. Heat flux in NSFs with different source/drain.

In summary, based on the previous analysis, the pure Si source/drain can provide the least self-heating (lowest R_{EFF}), but $Si_{0.5}Ge_{0.5}$ provides the optimal trade-off between the performance (I_{MAX}) and the self-heating (R_{EFF}).

4.4.10 Iso-T_{MAX} Performance Comparison

Based on the previous study, we identify that there are two design parameters that can reduce the self-heating in NSFs: 1) a smaller nanosheet spacing (T_{SUS}) (section 4.4.2), and 2) a smaller spacer length (section 4.4.5 and section 4.4.7). Although it is found that a larger sheet width (assuming the same difference between the sheet pitch and the sheet width) can relieve the self-heating, this cannot be a design parameter as the sheet width is fixed by circuit design considerations.

In this section, to investigate the best-case scenario for the NSF, we add the two cases to the comparison in section 4.3.3: 1) L_G=12nm, L_{SP}=6nm, T_{SUS}=5nm and 2) L_G=14nm, L_{SP}=5nm, T_{SUS}=5nm. Note that a 5nm T_{SUS} is practically challenging to achieve since the gate oxide (oxide interfacial layer and high-k dielectric layer) from the adjacent nanosheets occupies 2×(1.28+0.4)nm=3.36nm, leaving only 1.64nm (vertical thickness) for the entire gate workfunction metal stack and (possibly) the tungsten filling.

Subsequently, simulations were performed with a range of V_{DD} values to find the value for which T_{MAX} in NSF matches that in FF with $|V_{DD}|$=0.75V. A summary comparison is provided in Table 4.4-7.

	$\|V_{DD}\|$ (V)	I_{MAX} (mA/μm)	T_{MAX} (K)
N-FF	0.75	1.96 (100%)	393
P-FF	0.75	1.76 (100%)	396
N-GAAF	0.71	1.83 (93%)	385
P-GAAF	0.71	1.57 (89%)	396
N-NSF (Design 1)	0.72	1.76 (90%)	381
P-NSF (Design 1)	0.72	1.72 (98%)	397
N-NSF (Design 2)	0.73	1.94 (99%)	388
P-NSF (Design 2)	0.73	1.76 (100%)	396
N-NSF (Design 3)	0.71	1.94 (99%)	384
P-NSF (Design 3)	0.71	1.85 (105%)	397

Table 4.4-7 V_{DD} scaling of NSFs and GAAFs to match T_{MAX} of FFs. For NSF, design 1 is the reference case in section 4.3.3 (L_G=12nm, L_{SP}=6nm, and T_{SUS}=9nm). Design 2 features L_G=12nm, L_{SP}=6nm, and T_{SUS}=5nm. Design 3 features L_G=14nm, L_{SP}=5nm, and T_{SUS}=5nm. The percentages in parentheses are the current ratios to the corresponding FF.

It is found that both design 2 and design 3 of NSFs can achieve similar performance to that of FF by reducing the supply voltage to achieve the same T_{MAX} in p-channel transistors. It should be noted that although design 3 has larger (+5%) I_{ON} than that of design 2, its total gate capacitance increases by 21% (141aF@0.73V in design 2 vs. 170aF@0.71V), which is due to a thinner gate sidewall spacer. Therefore, design 2 still has performance advantage and is considered the best NSF case.

4.5 Conclusion

In section 4.3, the performance of FinFETs (FFs), Nanosheet FETs (NSFs), and Gate-All-Around FETs (GAAFs) are evaluated and compared in the context of self-heating. It is found that with similar on-state current, FFs have the least

self-heating (as indicated by the lowest T_{MAX}), while GAAFs are the worst. And in general, n-channel transistors have less self-heating than their p-channel counterparts. In order for NSFs and GAAFs to achieve the same T_{MAX}, their supply voltages must be lowered. Under these voltages, NSFs and GAAFs lose their performance advantage to FFs.

In section 4.4, a new metric "specific thermal resistance" (R_{EFF}) is defined to gauge SHE for a fair comparison between different transistor structures This metric normalizes the maximum temperature rise to the layout area and power generated by Joule heating. The electro-thermal simulation is performed to evaluate the SHE when p-channel NSF design parameters (e.g., nanosheet spacing, raised source/drain height, gate sidewall spacer length, source/drain Ge molefraction) are varied. Effects of Fin Pitch and Sheet Pitch on SHE were also studied. The pure thermal simulation, which assumes a constant heat power source, is also performed in relevant cases and it enables much shorter turn-around time. In conclusion, it is found that:

- The heat dissipates primarily downwards to the substrate. In n-channel transistors, the heat propagates through the silicon source/drain to the substrate. In p-channel transistors, however, due to a much smaller thermal conductivity of $Si_{1-x}Ge_x$, most of the heat is dissipated through the bottom spacers to the substrate.
- A smaller nanosheet spacing results in less SHE.
- The raised source/drain height (i.e., the height over the top conductive channel surface) has minimal impact on SHE.
- When the sheet width is fixed, a larger sheet pitch worsens the SHE due to the waste of area. When the difference between the sheet pitch and the sheet width is fixed, a larger sheet width helps reduce SHE, as it increases the substrate to STI area ratio.
- A thinner gate sidewall spacer (i.e., smaller L_{SP}) results in less SHE.
- The source/drain length has minimal impact on SHE.
- SHE is more severe when the source/drain material composition deviates from pure Si or Ge.
- FF has the least SHE of all the advanced FET structures. The NSF is relatively more susceptible to SHE, compared to FF.

The optimal NSF design (with 6nm L_{SP}, 12nm L_{SD} and 5nm T_{SUS}) can achieve similar performance to that of the FF under the same T_{MAX} constraint (*i.e.* operating the NSF with lower V_{DD}).

4.6 Appendix: Validity of Using a 1-Fin FF/GAAF with AreaFactor=4 to Simulate 4-Fin FF/GAAF

In this section, we compare a full multi-fin FF/GAAF simulation and a single-fin FF/GAAF simulation with appropriate scaling factors to examine the correctness of the approach used in this chapter.

Figure 4.6-1 shows the full structure of the 4-fin FF and GAAF. For clarity, the source region is omitted. All the fins share the same substrate contact at the bottom (not shown in the plot).

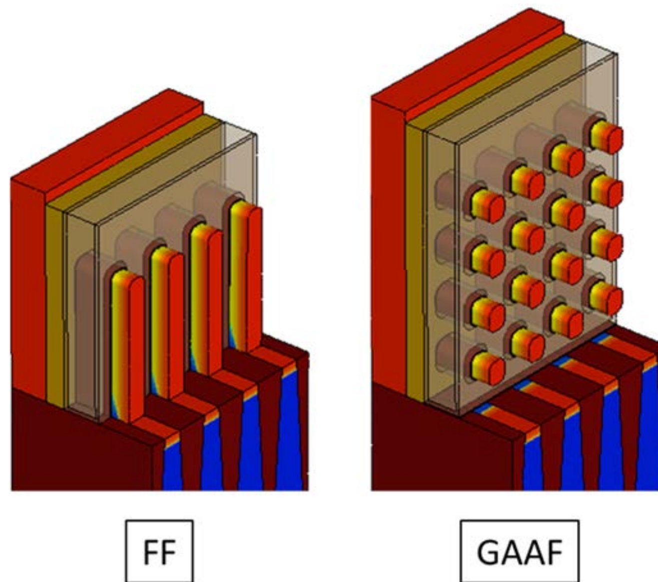


Figure 4.6-1. A full 4-fin FF and GAAF used for comparison.

As shown in Figure 4.6-2, the simulated electrical characteristics and the thermal characteristics of FFs are very similar. The current of the 1-fin case is scaled by a factor of 4 to account for the fact that a 4-fin structure is intended to be used. The results for GAAFs are similar (not shown). Figure 4.6-3 further shows that the gate capacitance are also the same under these two scenarios.

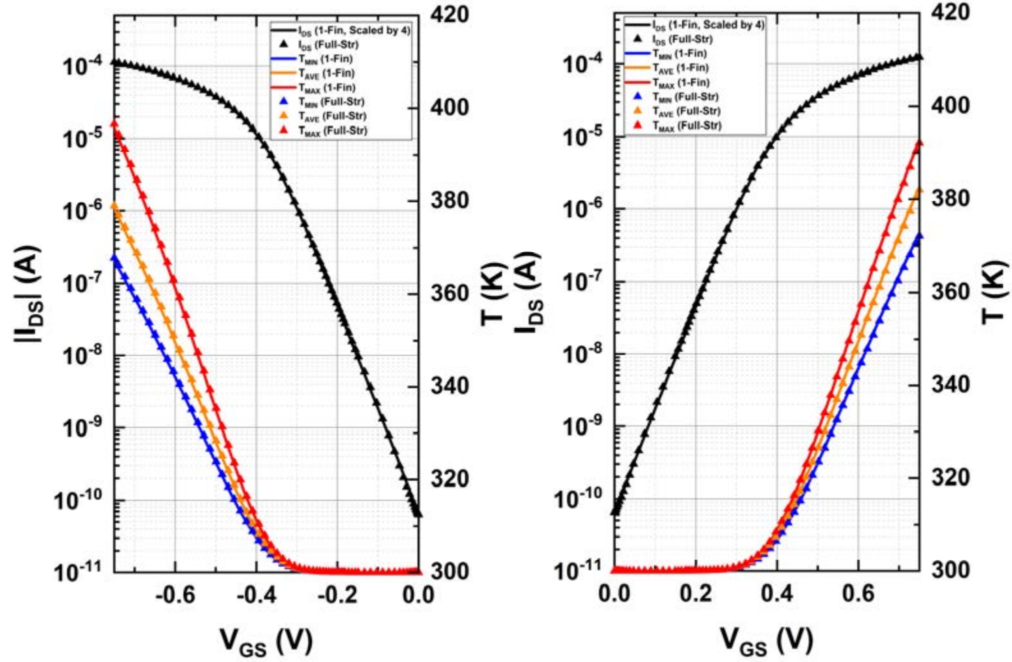


Figure 4.6-2. Comparison of simulated I_{DS}-V_{GS} and temperatures between using a 1-fin FF with 4X scaling and using a full 4-fin FF.

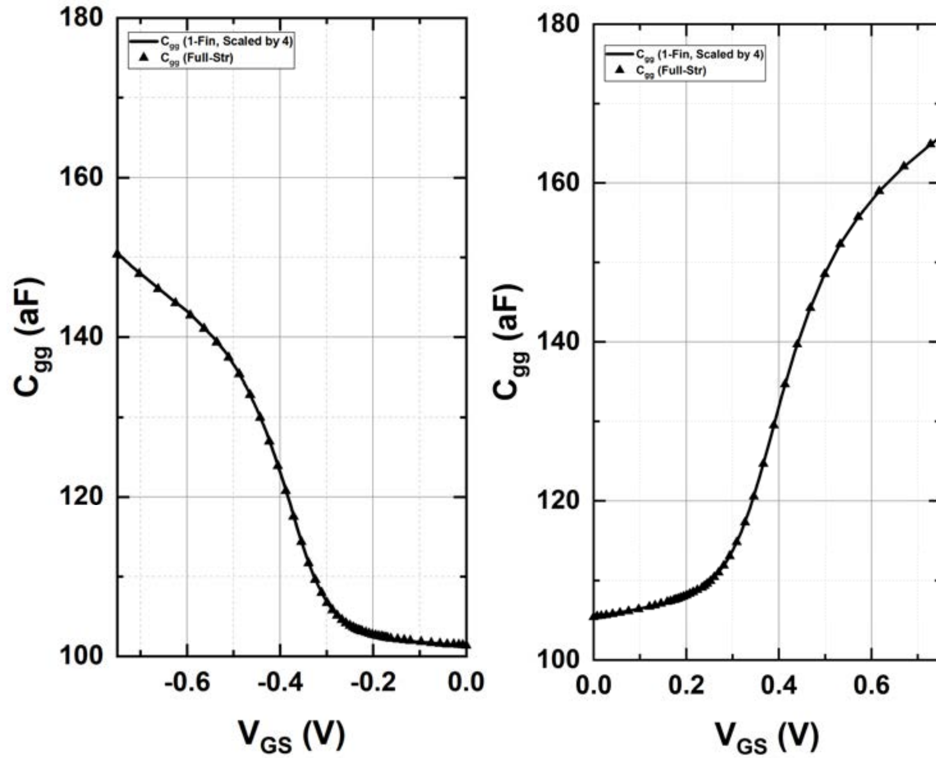


Figure 4.6-3. Comparison of simulated C_{gg}-V_{GS} between using a 1-fin FF with 4X scaling and using a full 4-fin FF.

Due to a significant reduction of meshes required, the computation time has been observed to reduce by at least 4X; in some cases, the improvement is >10X. Therefore, by using a single fin FF/GAAF with appropriate scaling factor (in this case, 4), the simulation can correctly capture the characteristics of the full structure with much shorter turn-around time.

4.7 References

[1] S. L. Liu, J. J. Horng, Amit Akundu, Y. C. Hsu, B. S. Lien, S. F. Liu, C. W. Chang, H. D. Hsieh, D. S. Huang, Y. C. Peng, S. Liu and M. Chen, "Self-Heating Temperature Behavior Analysis for DC - GHz Design Optimization in Advanced FinFETs," 2019 Symposium on VLSI Technology, Kyoto, Japan, 2019, pp. T200-T201.

[2] J. Choi, U. Monga, Y. Park, H. Shim, U. Kwon, S. Pae and D. S. Kim, "Impact of BEOL Design on Self-heating and Reliability in Highly-scaled FinFETs," 2019 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), Udine, Italy, 2019, pp. 1-4.

[3] R. Chau, "Challenges and opportunities of emerging nanotechnology for VLSI nanoelectronics," 2007 International Semiconductor Device Research Symposium, College Park, MD, 2007, pp. 1-1.

[4] (2019) Samsung Foundry Forum. [Online]. Available: https://news.samsung.com/global/samsung-electronics-leadership-in-advanced-foundry-technology-showcased-with-latest-silicon-innovations-and-ecosystem-platform.

[5] D. Jang, D. Yakimets, G. Eneman, P. Schuddinck, M. G. Bardon, P. Raghavan, A. Spessot, D. Verkest and A. Mocuta, "Device Exploration of NanoSheet Transistors for Sub-7-nm Technology Node," in IEEE Transactions on Electron Devices, vol. 64, no. 6, pp. 2707-2713, June 2017.

[6] G. Bae, D.-I. Bae, M. Kang, S.M. Hwang, S.S. Kim, B. Seo, T.Y. Kwon, T.J. Lee, C. Moon, Y.M. Choi, K. Oikawa, S. Masuoka, K.Y. Chun, S.H. Park, H.J. Shin, J.C. Kim, K.K. Bhuwalka, D.H. Kim, W.J. Kim, J. Yoo, H.Y. Jeon, M.S. Yang, S.-J. Chung, D. Kim, B.H. Ham, K.J. Park, W.D. Kim, S.H. Park, G. Song and Y.H. Kim, "3nm GAA Technology featuring Multi-Bridge-Channel FET for Low Power and High Performance Applications," 2018 IEEE

International Electron Devices Meeting (IEDM), San Francisco, CA, 2018, pp. 28.7.1-28.7.4.

[7] J. Ryckaert, M. H. Na, P. Weckx, D. Jang, P. Schuddinck, B. Chehab, S. Patli, S. Sarkar, O. Zografos, R. Baert and D. Verkest, "Enabling Sub-5nm CMOS Technology Scaling Thinner and Taller!," 2019 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2019, pp. 29.4.1-29.4.4.

[8] L. Cai, W. Chen, G. Du, J. Kang, X. Zhang and X. Liu, "Investigation of self-heating effect on stacked nanosheet GAA transistors," 2018 International Symposium on VLSI Technology, Systems and Application (VLSI-TSA), Hsinchu, 2018, pp. 1-2.

[9] W. Chen, L. Cai, K. Wang, X. Zhang, X. Liu and G. Du, "Self-heating induced Variability and Reliability in Nanosheet-FETs Based SRAM," 2018 IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA), Singapore, 2018, pp. 1-4.

[10] M. J. Kang, I. Myeong, M. Kang and H. Shin, "Analysis of DC Self Heating Effect in Stacked Nanosheet Gate-All-Around Transistor," 2018 IEEE 2nd Electron Devices Technology and Manufacturing Conference (EDTM), Kobe, 2018, pp. 343-345.

[11] Sentaurus User’s Manual, Version L-2016.03, Synopsys, Inc., Mountain View, CA, USA.

[12] N. Loubet, T. Hook, P. Montanini, C.-W. Yeung, S. Kanakasabapathy, M. Guillom, T. Yamashita, J. Zhang, X. Miao, J. Wang, A. Young, R. Chao, M. Kang, Z. Liu, S. Fan, B. Hamieh, S. Sieg, Y. Mignot, W. Xu, S.-C. Seo, J. Yoo, S. Mochizuki, M. Sankarapandian, O. Kwon, A. Carr, A. Greene, Y. Park, J. Frougier, R. Galatage and R. Bao, "Stacked nanosheet gate-all-around transistor to enable scaling beyond FinFET," 2017 Symposium on VLSI Technology, Kyoto, 2017, pp. T230-T231.

[13] S. A. Mujtaba, Advanced Mobility Models for Design and Simulation of Deep Submicrometer MOSFETs, Ph.D. thesis, Stanford University, Stanford, CA, USA, December 1995.

[14] S. Reggiani, E. Gnani, A. Gnudi, M. Rudan and G. Baccarani, "Low-Field Electron Mobility Model for Ultrathin-Body SOI and Double-Gate MOSFETs

With Extremely Small Silicon Thicknesses," IEEE Transactions on Electron Devices, vol. 54, no. 9, pp. 2204-2212, 2007.

[15] M. G. Ancona and G. J. Iafrate, "Quantum correction to the equation of state of an electron gas in a semiconductor," Physical Review B, vol. 39, no. 13, pp. 9536-9540, 1989.

[16] M. G. Ancona, "Density-gradient theory: A macroscopic approach to quantum confinement and tunneling in semiconductor devices," J. Comput. Electron., vol. 10, nos. 1–2, pp. 65–97, 2011.

[17] G. Wachutka, "An Extended Thermodynamic Model for the Simultaneous Simulation of the Thermal and Electrical Behaviour of Semiconductor Devices", Proceedings of the Sixth International Conference on the Numerical Analysis of Semiconductor Devices and Integrated Circuits (NASECODE VI), Dublin, Ireland, pp. 409-414, July 1989.

[18] V. Vidya, "Thin-Body Silicon FET Devices and Technology," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Univ. California Berkeley, Berkeley, CA, USA, 2007.

[19] F. Schaffler, "Silicon-Germanium (Si_1-xGe_x)," in Properties of Advanced Semiconductor Materials GaN, AlN, InN, BN, SiC, SiGe, M. E. Levinshtein, S. L. Rumyantsev, M. S. Shur, Eds., New York: John Wiley & Sons, Inc., 2001, 149-188.

[20] A. Delan, M. Rennau, S. E. Schulz, T. Gessner, "Thermal conductivity of ultra low-k dielectrics," Center for Microtechnologies, Chemnitz University of Technology, D-09107 Chemnitz, Germany. doi: 10.1016/S0167-9317(03)00417-9.

[21] M. Wagner, "Simulation of Thermoelectric Devices," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Technischen Universität Wien, Vienna, Austria, 2007. [Online]. Available: https://www.iue.tuwien.ac.at/phd/mwagner.

Chapter 5

Conclusion

5.1 Summary and Contributions of This Work

Moore's law predicts a doubling count of transistors per chip roughly every two years [1], which usually translates into halving of the layout area per transistor in practice since an increase in chip size is limited by cost considerations [2]. The industry has been successful in following Moore's law during the past five decades, even after Dennard scaling [3] became ineffective around 2005. However, as the industry heads down to single-digit nanometer technology nodes [4, 5], transistor area scaling becomes even more difficult due to ever increasing cost as a result of more complex fabrication process (Figure 5.1-1).

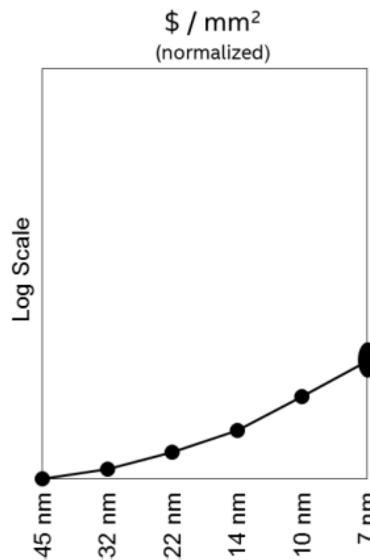


Figure 5.1-1. Costs per area are increasing dramatically. Adapted from [6].

This dissertation aims to recognize potential problems and propose possible solutions to these problems as we continue to shrink the area of transistors in different ways.

To increase the current density of a FinFET, high mobility channel materials eventually will be required because the approach of increasing fin height starts to

meet fabrication challenges and diminishing returns in performance gain. Low Ge molefraction silicon germanium (SiGe) is considered one of the most promising candidate channel materials for p-channel FinFETs. However, due to increased permittivity and lower bandgap as compared to silicon, the SiGe FinFET is expected to have worse electrostatic integrity than Si FinFET. To mitigate this problem, in **chapter 2**, the hetero-channel FinFET is proposed. The hetero-channel region comprises a heterogeneous layer sandwich of silicon (Si) and silicon-germanium (i.e., $Si_{0.9}Ge_{0.1}$). Due to the higher bandgap region (Si) near the source and drain regions, the hetero-channel design can achieve comparable electrostatic integrity as the conventional Si FinFET. Moreover, due to the higher mobility induced by stress in $Si_{0.9}Ge_{0.1}$, the hetero-channel design provides for +16% larger and +9% on-state current compared to Si FinFET and $Si_{0.9}Ge_{0.1}$ FinFET, respectively. It should be noted that only 10% Ge is used; hence a complex process such as aspect ratio trapping (ART) method is not required, for ease of manufacture. Therefore, the hetero-channel FinFET is promising for future low-power applications. Larger Ge molefraction SiGe might achieve even higher performance, but the increased sensitivity to the recess length makes them less favorable than $Si_{0.9}Ge_{0.1}$.

6-T SRAM is one of the most often utilized circuit block in modern SoCs and for cache memory in CPUs. One particular bit cell design - high density cell (HDC) is of focus since it is the most compact (i.e., occupying the smallest layout area) among all bit cell designs. Due to the large number of 6-T SRAM circuits, voltage scaling becomes critical for power reduction. However, due to the small geometry, the HDC design usually imposes more stringent requirements on transistor process variations and design robustness. For an advanced transistor technology, in order for the HDC to function properly at low voltages, circuit-level assist techniques are generally required, which results in larger chip area and greater power consumption. In **chapter 3**, a novel scheme for controllably reducing the drive strength of iFinFET (inserted-oxide FinFET) is proposed to facilitate voltage scaling of the 6-T SRAM HDC cell. Specifically, one or more nanowire channels in iFinFETs are selectively doped so that the threshold voltage of the doped nanowires is increased so that, under normal operating conditions, these doped nanowires do not conduct current. It is shown that after taken the systematic and random sources of variations into consideration, the best design can enable more than 0.1V reduction in the minimum cell operating voltage (V_{MIN}).

As a side effect of aggressive transistor miniaturization, self-heating effects become more severe. This effect limits transistor performance gains because of degraded mobility and transistor reliability. In the first half of **chapter 4**, three promising candidate advanced transistor structures, namely, FinFETs (FFs), Nanosheet FETs (NSFs), and Gate-All-Around Nanowire FETs (GAAFs) are compared in terms of performance under the constraint of maintaining the same peak temperature. It is found that n-channel transistors have much less self-heating than their p-channel counterparts. This is due to the fact that in n-channel transistors, the majority of heat can be dissipated through the silicon source/drain regions. Differences in SHE for n-channel transistor structures are not significant. On the other hand, due to the poorer thermal conductivity (10X less than that of silicon) of silicon germanium, heat cannot be well dissipated through the source/drain regions in p-channel transistor structures. In p-channel FFs, the heat can still propagate through the fin to the substrate; hence FFs have the least self-heating effects among all the advanced p-channel transistor structures. In GAAFs and NSFs, the Si channels (i.e., nanowires or nanosheets) are suspended from the silicon substrate. Hence the majority of the heat is dissipated through the bottom spacers, which are also poor thermal conductors. P-channel GAAFs have the worst self-heating since to achieve a similar on-state current, more nanowires must be used; therefore, the localized heating in middle conductive channels are more severe than that in NSFs. When the operating (supply) voltage of GAAFs and NSFs is lowered so that their p-channel transistors have the same self-heating as p-channel FFs, to avoid worsening reliability, it is found that GAAFs and NSFs have worse performance than that of FFs, especially for n-channel transistors (n-channel and p-channel transistors share the same operating voltage).

To understand how various design parameters can impact NSFs self-heating, in the second half of **chapter 4**, a new metric "specific thermal resistance" (R_{EFF}) is introduced to gauge SHE for a fair comparison among different transistor structures and different design parameters. This metric normalizes the transistor maximum temperatures with respect to the layout area and the power generated by Joule heating. A smaller R_{EFF} is considered better (less self-heating). In this study, several design parameters are varied: nanosheet spacing, source/drain height, gate sidewall spacer length, source/drain Ge molefraction, sheet pitch, etc. In summary, it is found that:

- In p-channel NSFs, even though the source/drain regions are poor thermal conductors, the majority of heat is still dissipated to the

substrate heat sink instead of the upper gate or source/drain metal contacts. As a result, the raised source/drain height (i.e., the height over the top conductive channel surface) has minimal impact on SHE.

- A smaller nanosheet spacing is beneficial for SHE. But it may lead to larger process complexity.
- A larger sheet pitch worsens the SHE due to the waste of area as the sheet width is fixed.
- A larger sheet width helps reduce SHE when the difference between the sheet pitch and the sheet width is fixed. This is because it increases the substrate (heat sink) to STI area ratio.
- SHE is mitigated with a thinner gate sidewall spacer (i.e., smaller L_{SP}).
- The source/drain length has negligible impact on SHE.
- SHE is more severe when the source/drain material composition deviates from pure Si or Ge. But using pure Si or Ge source/drain can cause large performance degradation.

In the end, the operating voltage of the optimally designed NSFs is lowered to reduce the self-heating to the same level of FFs. In this scenario, the optimal NSF design can achieve similar performance as can FFs.

5.2 Future Directions

In **chapter 2**, the low Ge molefraction $Si_{1-x}Ge_x$, where x is small ($x < 25\%$) is investigated. This channel material can be directly fabricated on top of the conventional silicon substrate without process issues. In the future, to further improve the mobility, high Ge molefraction SiGe or even pure Ge is projected to be used. This usually requires a strain relaxed buffer (SRB) intermediate layer to avoid the formation of crystal dislocations due to large lattice mismatch. The concept of hetero-channel, which takes advantage of a higher bandgap channel region near the source/drain region to improve electrostatic integrity, can be extended to this case.

In **chapter 3**, the concept of fine tuning SRAM cell ratio via counter doping the top nanowire(s) in iFinFET can be extended into other SRAM bit cell designs, such as high performance cell (HPC). In these designs, transistors may feature multiple fins, which gives extra tuning knobs. In addition, physical

removal of top nanowire(s) via selective etching might also be worth investigating as it could potentially reduce the variations due to RDF.

In **chapter 4**, the self-heating in advanced silicon transistors is evaluated and compared. This study can be further extended into more advanced architectures (e.g., Fork-Sheet FETs [7]) and other channel materials (e.g., SiGe [4]). In addition, the performance evaluation can also be assessed dynamically in a circuit setup (e.g., inverter, multi-input NAND/NOR, ring oscillator).

5.3 References

[1] G. E. Moore, "Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff.," in IEEE Solid-State Circuits Society Newsletter, vol. 11, no. 3, pp. 33-35, Sept. 2006. doi: 10.1109/N-SSC.2006.4785860.

[2] M. Bohr, "A 30 Year Retrospective on Dennard's MOSFET Scaling Paper," in IEEE Solid-State Circuits Society Newsletter, vol. 12, no. 1, pp. 11-13, Winter 2007. doi: 10.1109/N-SSC.2007.4785534.

[3] R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous and A. R. Leblanc, "Design Of Ion-implanted MOSFET's with Very Small Physical Dimensions," in Proceedings of the IEEE, vol. 87, no. 4, pp. 668-678, April 1999. doi: 10.1109/JPROC.1999.752522.

[4] G. Yeap, S. S. Lin, Y. M. Chen, H. L. Shang, P. W. Wang, H. C. Lin, Y. C. Peng, J. Y. Sheu, M. Wang, X. Chen, B. R. Yang, C. P. Lin, F. C. Yang, Y. K. Leung, D. W. Lin, C. P. Chen, K. F. Yu, D. H. Chen, C. Y. Chang, H. K. Chen, P. Hung, C. S. Hou, Y. K. Cheng, J. Chang, L. Yuan, C. K. Lin, C. C. Chen, Y. C. Yeo, M. H. Tsai, H. T. Lin, C. O. Chui, K. B. Huang, W. Chang, H. J. Lin, K. W. Chen, R. Chen, S. H. Sun, Q. Fu, H. T. Yang, H. T. Chiang, C. C. Yeh, T. L. Lee, C. H. Wang, S. L. Shue, C. W. Wu, R. Lu, W. R. Lin, J. Wu, F. Lai, Y. H. Wu, B. Z. Tien, Y. C. Huang, L. C. Lu, J. He, Y. Ku, J. Lin, M. Cao, T. S. Chang and S. M. Jang, "5nm CMOS Production Technology Platform featuring full-fledged EUV, and High Mobility Channel FinFETs with densest 0.021µm2 SRAM cells for Mobile SoC and High Performance Computing Applications," 2019 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2019, pp. 36.7.1-36.7.4. doi: 10.1109/IEDM19573.2019.8993577.

[5] G. Bae, D.-I. Bae, M. Kang, S.M. Hwang, S.S. Kim, B. Seo, T.Y. Kwon, T.J. Lee, C. Moon, Y.M. Choi, K. Oikawa, S. Masuoka, K.Y. Chun, S.H. Park, H.J. Shin, J.C. Kim, K.K. Bhuwalka, D.H. Kim, W.J. Kim, J. Yoo, H.Y. Jeon, M.S. Yang, S.-J. Chung, D. Kim, B.H. Ham, K.J. Park, W.D. Kim, S.H. Park, G. Song and Y.H. Kim, "3nm GAA Technology featuring Multi-Bridge-Channel FET for Low Power and High Performance Applications," 2018 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2018, pp. 28.7.1-28.7.4.

[6] M. T. Bohr, "Logic Technology Scaling to Continue Moore's Law," 2018 IEEE 2nd Electron Devices Technology and Manufacturing Conference (EDTM), Kobe, 2018, pp. 1-3. doi: 10.1109/EDTM.2018.8421433.

[7] P. Weckx, J. Ryckaert, E. D. Litta, D. Yakimets, P. Matagne, P. Schuddinck, D. Jang, B. Chehab, R. Baert, M. Gupta, Y. Oniki, L. -A. Ragnarsson, N. Horiguchi, A. Spessot and D. Verkest, "Novel forksheet device architecture as ultimate logic scaling device towards 2nm," 2019 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2019, pp. 36.5.1-36.5.4.