-To-

My parents
ACKNOWLEDGMENTS

I would like to express my sincere appreciation to the chairman of my supervisory committee, Professor Jerry G. Fossum, for his guidance and support throughout the course of this work. His great knowledge in semiconductor physics motivated my devotion to the field of semiconductor devices. He was a role model for me, put things in proper perspective, and contributed to my positive attitude. I would also like to thank the members of my supervisory committee (Professors Robert Fox, Jing Guo, and Susan Sinnott) for their guidance and interest in this work.

I am grateful to Samsung Electronics, Freescale Semiconductor, and the National Science Foundation for their financial support and the useful information. I would also like to thank fellow students Ji-Woon Yang, Vishal Trivedi, Weimin Zhang, Murshed Chowdhury, Zhichao Lu, Siddharth Chouksey, and Shishir Agarwal for their insightful and technical discussions and friendships. Also, I thank all of the friends who made my years at the University of Florida such an enjoyable chapter of my life.

I must express heartfelt thanks to my mother (Kyung-Lim Lee), my sisters (Eun-Kyung, Yeon-Jung, and Hee-Jung), and my brother-in-law (Dong-Ju Park) for their constant encouragement, support, help, sacrifice, and love throughout my studies.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACKNOWLEDGEMENT</td>
<td>iii</td>
</tr>
<tr>
<td></td>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td></td>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td></td>
<td>KEY TO ABBREVIATIONS</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>ABSTRACT</td>
<td>xii</td>
</tr>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>NONCLASSICAL CMOS: POTENTIAL NONCLASSICAL TECHNOLOGIES VERSUS A HYPOTHETICAL BULK-SILICON TECHNOLOGY</td>
<td>8</td>
</tr>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>8</td>
</tr>
<tr>
<td>2.2</td>
<td>UFDG, UFPDB, and Simulation Conditions</td>
<td>10</td>
</tr>
<tr>
<td>2.3</td>
<td>Immunity to Short-Channel Effects</td>
<td>12</td>
</tr>
<tr>
<td>2.4</td>
<td>28nm Device Design</td>
<td>14</td>
</tr>
<tr>
<td>2.5</td>
<td>CMOS Performance Projections</td>
<td>22</td>
</tr>
<tr>
<td>2.6</td>
<td>Thin-BOX FD/SOI CMOS</td>
<td>32</td>
</tr>
<tr>
<td>2.7</td>
<td>Summary</td>
<td>35</td>
</tr>
<tr>
<td>3</td>
<td>BULK INVERSION IN FINFETS AND IMPLIED INSIGHTS ON EFFECTIVE GATE WIDTH</td>
<td>37</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>37</td>
</tr>
<tr>
<td>3.2</td>
<td>Numerical Simulations</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>3.2.1 I-V Characteristics of DG and TG FinFETs</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>3.2.2 Electric-Field Fringing Effects</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>3.2.3 Bulk Inversion</td>
<td>43</td>
</tr>
<tr>
<td>3.3</td>
<td>Implied Insight of Bulk Inversion</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>3.3.1 Proper Effective Gate Width</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>3.3.2 Layout Area</td>
<td>52</td>
</tr>
<tr>
<td>3.4</td>
<td>Summary</td>
<td>55</td>
</tr>
<tr>
<td>Table</td>
<td>page</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>5.1 UFDG/Spice3-predicted sensitivity to the variation of straggle, fin width, and the channel length</td>
<td>107</td>
<td></td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>UFDG-predicted threshold voltage roll-off and DIBL</td>
<td>13</td>
</tr>
<tr>
<td>2.2</td>
<td>UFDG-predicted current-gate voltage characteristics</td>
<td>16</td>
</tr>
<tr>
<td>2.3</td>
<td>UFDG-predicted effects of separately varying toxf and toxb</td>
<td>18</td>
</tr>
<tr>
<td>2.4</td>
<td>UFDG/UFPDB-predicted current-gate voltage characteristics</td>
<td>20</td>
</tr>
<tr>
<td>2.5</td>
<td>UFDG/ and UFPDB/Spice3-predicted propagation delays</td>
<td>23</td>
</tr>
<tr>
<td>2.6</td>
<td>UFDG/ and UFPDB/Spice3-predicted propagation delays</td>
<td>24</td>
</tr>
<tr>
<td>2.7</td>
<td>UFDG- and UFPDB-predicted gate capacitances</td>
<td>27</td>
</tr>
<tr>
<td>2.8</td>
<td>UFDG/Spice3-predicted loaded CMOS ring-oscillator delays</td>
<td>31</td>
</tr>
<tr>
<td>2.9</td>
<td>UFDG/Spice3-predicted propagation delays</td>
<td>34</td>
</tr>
<tr>
<td>3.1</td>
<td>Two-dimensional cross-section view of the DG FinFET</td>
<td>39</td>
</tr>
<tr>
<td>3.2</td>
<td>Davinci-predicted current-voltage characteristics</td>
<td>41</td>
</tr>
<tr>
<td>3.3</td>
<td>Davinci-predicted on-state current increase</td>
<td>42</td>
</tr>
<tr>
<td>3.4</td>
<td>Davinci-predicted on-state electron density along the top fin surface, at the center of the channel</td>
<td>44</td>
</tr>
<tr>
<td>3.5</td>
<td>Davinci-predicted on-state electron density down the middle of the fin, at the center of the channel</td>
<td>45</td>
</tr>
<tr>
<td>3.6</td>
<td>Medici-predicted electron density</td>
<td>49</td>
</tr>
<tr>
<td>3.7</td>
<td>Davinci-predicted on-state electron density at the middle of the top fin-body surface, at the center of the channel</td>
<td>51</td>
</tr>
<tr>
<td>3.8</td>
<td>Calculated gate layout-area ratios of TG and DG FinFETs</td>
<td>54</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>4.1</td>
<td>Medici-predicted low-frequency gate capacitance</td>
<td>59</td>
</tr>
<tr>
<td>4.2</td>
<td>Medici-predicted low-frequency subthreshold gate capacitance</td>
<td>62</td>
</tr>
<tr>
<td>4.3</td>
<td>A schematic diagram of the gate-source/drain structure</td>
<td>63</td>
</tr>
<tr>
<td>4.4</td>
<td>Basic two-plate model for the fringe capacitance</td>
<td>65</td>
</tr>
<tr>
<td>4.5</td>
<td>Schematics of the G-S/D underlap structure for the weak inversion analysis</td>
<td>66</td>
</tr>
<tr>
<td>4.6</td>
<td>Illustration of how varying the actual lateral doping density profile changes the weak inversion effective channel length</td>
<td>68</td>
</tr>
<tr>
<td>4.7</td>
<td>Schematic of the SG FD/SOI MOSFET with G-S/D underlap</td>
<td>71</td>
</tr>
<tr>
<td>4.8</td>
<td>UFDG- and Medici-predicted gate capacitance</td>
<td>77</td>
</tr>
<tr>
<td>4.9</td>
<td>UFDG- and Medici-predicted current-voltage characteristics</td>
<td>79</td>
</tr>
<tr>
<td>4.10</td>
<td>UFDG/Spice3-predicted propagation delays</td>
<td>81</td>
</tr>
<tr>
<td>4.11</td>
<td>UFDG- and Medici-predicted gate capacitance</td>
<td>82</td>
</tr>
<tr>
<td>5.1</td>
<td>UFDG-predicted threshold voltage</td>
<td>92</td>
</tr>
<tr>
<td>5.2</td>
<td>Threshold voltage reduction</td>
<td>94</td>
</tr>
<tr>
<td>5.3</td>
<td>UFDG-predicted on-state current and the increased source/drain series resistance</td>
<td>95</td>
</tr>
<tr>
<td>5.4</td>
<td>UFDG-predicted threshold voltage</td>
<td>97</td>
</tr>
<tr>
<td>5.5</td>
<td>UFDG-predicted read static noise margin versus the effective channel length and schematics of 6-T SRAM cell</td>
<td>99</td>
</tr>
<tr>
<td>5.6</td>
<td>UFDG-predicted butterfly curves</td>
<td>101</td>
</tr>
<tr>
<td>5.7</td>
<td>UFDG-predicted write-0 margin</td>
<td>102</td>
</tr>
<tr>
<td>5.8</td>
<td>UFDG/Spice3-predicted sensitivity of the optimal DG FinFET</td>
<td>105</td>
</tr>
<tr>
<td>5.9</td>
<td>UFDG-predicted read-SNM versus the physical gate length</td>
<td>109</td>
</tr>
<tr>
<td>5.10</td>
<td>UFDG-predicted butterfly curves</td>
<td>112</td>
</tr>
</tbody>
</table>
5.11 UFDG-predicted write-0 margin ........................................... 113
5.12 UFDG-predicted current-gate voltage characteristics ............... 114
5.13 UFDG-predicted gate capacitance........................................ 116
5.14 UFDG/Spice3-predicted propagation delays per stage............... 117
5.15 UFDG/Spice3-predicted loaded CMOS ring-oscillator delays ...... 119
5.16 UFDG/Spice3-predicted sensitivity comparison of the two underlapped DG FinFET designs to the variation of straggle............... 121
5.17 UFDG/Spice3-predicted sensitivity comparison of the two underlapped DG FinFET designs to the variation of straggle............... 122
A.1 Representative potential variations in y at a given x and corresponding linear approximations ................................................. 134
B.1 Medici-predicted gate capacitance at low and high VDS ............ 138
B.2 Medici-predicted saturation gate capacitance .......................... 140
B.3 Medici-predicted saturation gate capacitance .......................... 141
B.4 Medici-predicted electron density profile across the SOI film ...... 143
B.5 Integrated inversion charge and its differentiation ................. 144
B.6 Medici-predicted surface and bulk electron density and its differentiation ................................................................. 146
B.7 Equivalent circuits of a DG MOS capacitor in the strong inversion region ................................................................. 147
**KEY TO ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADG</td>
<td>asymmetrical double-gate</td>
</tr>
<tr>
<td>BOX</td>
<td>buried-oxide</td>
</tr>
<tr>
<td>CMOS</td>
<td>complementary metal-oxide-semiconductor</td>
</tr>
<tr>
<td>DG</td>
<td>double-gate</td>
</tr>
<tr>
<td>DIBL</td>
<td>drain-induced barrier lowering</td>
</tr>
<tr>
<td>EOT</td>
<td>equivalent oxide thickness</td>
</tr>
<tr>
<td>FD</td>
<td>fully depleted</td>
</tr>
<tr>
<td>GIDL</td>
<td>gate-induced-drain leakage</td>
</tr>
<tr>
<td>G-S/D</td>
<td>gate-source/drain</td>
</tr>
<tr>
<td>HP</td>
<td>high performance</td>
</tr>
<tr>
<td>LOP</td>
<td>low operating power</td>
</tr>
<tr>
<td>LSTP</td>
<td>low standby power</td>
</tr>
<tr>
<td>MOSFET</td>
<td>metal-oxide-semiconductor field-effect transistor</td>
</tr>
<tr>
<td>PD</td>
<td>partially depleted</td>
</tr>
<tr>
<td>QM</td>
<td>quantum mechanical</td>
</tr>
<tr>
<td>RO</td>
<td>ring oscillator</td>
</tr>
<tr>
<td>SCE</td>
<td>short-channel effect</td>
</tr>
<tr>
<td>S/D</td>
<td>source/drain</td>
</tr>
<tr>
<td>SDG</td>
<td>symmetrical double-gate</td>
</tr>
<tr>
<td>SG</td>
<td>single gate</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>SNM</td>
<td>static noise margin</td>
</tr>
<tr>
<td>SOI</td>
<td>silicon-on-insulator</td>
</tr>
<tr>
<td>SRAM</td>
<td>static random access memory</td>
</tr>
<tr>
<td>TG</td>
<td>triple-gate</td>
</tr>
<tr>
<td>UFDG</td>
<td>University of Florida double-gate (model)</td>
</tr>
<tr>
<td>UFPDB</td>
<td>University of Florida partially depleted SOI and bulk MOSFET (model)</td>
</tr>
<tr>
<td>UTB</td>
<td>ultra-thin body</td>
</tr>
</tbody>
</table>
This dissertation addresses performance projections, design optimization, and physical modeling issues of nonclassical nanoscale CMOS devices with UTBs, assessing their potential to become the basis of the near-future mainstream semiconductor technology.

With regard to speed and immunity to SCEs, DG MOSFETs are projected to be generally superior to the SG counterparts because of their better gate control and higher drive currents. However, for light loads and moderate supply voltages, a suboptimal SG FD/SOI MOSFET design for both LOP and HP CMOS applications is found to yield speeds comparable to the DG designs based on their much lower intrinsic $C_G$, even though its current drive is much lower and its SCEs are much more severe. Compared to nonclassical CMOS designs, the delay of SG bulk-Si CMOS is predicted to be much longer due mainly to its high $C_G$ in the weak/moderate inversion region and relatively low drive current.
Relative values of $I_{on}$ in undoped-UTB DG and TG FinFETs are examined via 3-D numerical device simulations. The simulation results reveal significant bulk inversion in the fin bodies, which limits the benefit of the third (top) gate in the TG FinFET and which negates the utility of the commonly defined effective gate width ($W_{eff} = 2h_{Si} + w_{Si}$). Even the concept of $W_{eff}$ for the TG FinFET is invalidated, but the proper $W_{eff}$ for the DG FinFET is defined. Physical insights attained from the simulations further solidify our notion, based previously on gate layout-area inefficiency, that the third gate is neither desirable nor beneficial.

Parasitic G-S/D fringe capacitance in nonclassical nanoscale CMOS devices is shown, using 2-D numerical simulations, to be very significant, gate bias-dependent, and substantially reduced by well designed G-S/D underlap. Analytical modeling of the outer and inner components of the fringe capacitance is developed and verified by the numerical simulations; a BOX-fringe component is modeled for SG FD/SOI MOSFETs. With the new modeling implemented in UFDG, UFDG/Spice3 shows how nanoscale DG CMOS speed is severely affected by the fringe capacitance, and how this effect can be moderated by optimal underlap.

Based on the trade-off between SCEs and $I_{on}$, an optimal underlap, which is defined by short $L_{ext}$ and $\sigma_L$, is defined for SRAM applications. This optimization gives high $V_t$ along with small loss of $I_{on}$. For the CMOS speed issue, devices are optimally designed with long $L_{ext}$ and $\sigma_L$ since the latter tends to decrease $V_t$ and thus keep $I_{on}$ high, while the parasitic capacitance in weak inversion decreases with the underlap. With regard to the sensitivity issue, relatively long underlap, via short $L_{ext}$ and $\sigma_L$, is generally beneficial for both the SRAM applications and the CMOS speed.
Scale-down of device dimensions in conventional bulk-silicon CMOS technology has been a primary driving force of the semiconductor industry development over the past three decades. The better performance with the smaller size of the devices has been the basis of this development. However, for conventional bulk-Si (and partially depleted (PD) SOI) CMOS, continued scaling much beyond a physical gate length ($L_g$) of ~50nm [Sem01] is doubtful. This is because of severe short-channel effects (SCEs), high off-state leakage currents, and unacceptably low $I_{on}/I_{off}$ ratios. Indeed, controlling the body doping within very small dimensions, which is required for SCE control, has been the most difficult technological challenge to overcome for further scaling. Hence, there is a growing interest in nonclassical fully depleted (FD) SOI single-gate (SG) and double-gate (DG) MOSFETs with ultra-thin bodies (UTBs), which have inherent suppression of SCEs. Their small intrinsic gate capacitance in weak/moderate inversion and, especially for DG devices, the high $I_{on}/I_{off}$ ratio stemming from the nearly ideal subthreshold gate swing imply substantial CMOS speed superiority over the classical SG counterparts [Fos02]. However, DG technology is complex; the DG FinFET [His98, Hua99] is easiest to fabricate, but its proven utility is years away.
Contrarily, FD/SOI SG technology is less complicated; SOI UTBs and metal gates are the main obstacles in its development [Cho00].

Because of the technological complexities and difficulties associated with DG CMOS, questions have been posed about the performance advantage, relative to SG CMOS, that it can potentially provide. For example, if the DG MOSFET gives twice the current, but with twice the gate capacitance, then excessive device parasitics implied by the complex technology might render inferior performance. Further, it has been argued that SCEs in the bulk-Si SG MOSFET could be effectively suppressed by super-halo channel doping such that bulk-Si CMOS could actually be scaled down to 25nm channel lengths [Tau98]. However, this argument is simulation-based, and there is uncertainty about the physical modeling assumed [Ge01] and whether the assumed device structure could even be fabricated [Tau98]. Nonetheless, given such a “hypothetical” nanoscale bulk-Si CMOS technology, more detailed insights on the relative performance potentials of nonclassical UTB CMOS would be useful in deciding how and if they should be aggressively pursued.

In chapter 2, using our process/physics-based compact models (UFDG [Fos03a] and UFPDB [Fos97]) in Spice3, we project device characteristics and CMOS performances of nonclassical UTB CMOS technologies (FD/SOI and DG) and classical, hypothetical bulk-Si technologies optimized at the \(L_g = 28\text{nm}\) node. Comparisons of predicted SCEs of nonclassical devices and speed (RO delays) of the nonclassical and classical CMOS technologies are made, and good physical insights
regarding their relative characteristics are given. Namely, we first compare asymmetrical and symmetrical DG, and FD/SOI SG devices via simulations done with our process/physics-based compact model UFDG [Fos03a, Fos04a], with emphasis on their immunity to SCEs. Then, using UFDG, we optimally design these nonclassical devices for 28nm gate length, and project their characteristics, including CMOS ring-oscillator delays, which we compare with projections of a hypothetical 28nm bulk-Si SG CMOS design derived, using our UFPDB compact model [Fos97], from Taur et al. [Tau98]. Interpretations of the simulation results give good physical insights on the nonclassical technologies, and indicate which ones might best replace the classical technologies at nanoscale nodes of the SIA ITRS [Sem01].

While the DG FinFET has become a leading device option for future nanoscale CMOS, there is a technological limit to the aspect ratio (Rf) of the Si-fin height (h_{Si}) to the width (w_{Si}). Since w_{Si} must be ultra-thin for good control of SCEs [Fos04b], this limit implies small effective gate width (commonly assumed to be W_{eff} \equiv 2h_{Si}) and, ostensibly, low I_{on} per pitch. There is therefore interest in making the FinFET a triple-gate (TG) transistor by activating the top gate, yielding, from a surface inversion-charge perspective, W_{eff} \equiv 2h_{Si} + w_{Si} as is commonly assumed, and alleviating, with a doped fin-body, the thin-w_{Si} requirement [Doy03]. However, because of fin-corner effects [Fos03b] and technological limitations [Tri03a], the fin-body must be left undoped, and so relaxation of the thin-w_{Si} (i.e., UTB) requirement for SCE control for the TG FinFET,
relative to that for the DG device, is minimal [Fos04b]. Nonetheless, the larger \( W_{\text{eff}} \) for the TG device could mean significantly higher \( I_{\text{on}} \) relative to that of the DG FinFET even when \( w_{\text{Si}} \) is thin for good SCE control.

In chapter 3, to check the benefit of activating the top gate of FinFETs, relative values of \( I_{\text{on}} \) in undoped-body DG and TG FinFETs are examined via three-dimensional (3-D) numerical device simulations [Dav03]. Simulation results show that fin-body bulk inversion in strong inversion limits the benefit of the third (top) gate in the TG FinFET, and the commonly defined \( W_{\text{eff}} \) is inappropriate as an indicator of \( I_{\text{on}} \). Thus, we define the proper \( W_{\text{eff}} \) for the DG FinFET reflecting bulk inversion and, based on this proper (re)definition of \( W_{\text{eff}} \), examine the gate layout-area issue [Yan05] of FinFET CMOS.

Nonclassical nanoscale silicon CMOS devices, e.g., DG and SG FD/SOI MOSFETs with undoped UTBs, should be designed with gate-source/drain (G-S/D) underlap [Tri05a]. The benefits of the underlap include better control of SCEs via a gate bias-dependent effective channel length (\( L_{\text{eff}} \)) [Fos03c, Tri05a], as well as elimination of gate-induced drain leakage (GIDL) [Tan05] and gate-drain/source tunneling currents. The underlap, however, must be optimally designed because it tends to increase the source/drain (S/D) series resistance (\( R_{\text{SD}} \)) and decrease \( I_{\text{on}} \) [She03, Tan05].

Fringe capacitance in classical MOSFETs, with G-S/D overlap, was modeled some time ago [Shr82], and some modeling was recently
reported [Ban05] for DG MOSFETs with underlap. However, the modeling of Bansal et al. [Ban05] focused on a bias-independent outer-fringe capacitance, and did not address the $V_{GS}$ dependence of it as well as the inner-fringe component, which is quite important in nanoscale devices.

In chapter 4, we show, by device and circuit modeling and simulation, the significance and gate-bias dependence of parasitic fringe capacitance in nonclassical CMOS devices with G-S/D underlap. Based on the insights derived from numerical device simulations, we develop a complete analytical model for parasitic capacitance in nonclassical devices with G-S/D underlap, which includes both the outer- and inner-fringe components with $V_{GS}$ dependences, as well as a BOX-fringe component in the FD/SOI MOSFET. The new modeling is verified by 2-D numerical device simulations. Using a new version of UFDG with the parasitic fringe capacitance model, we check RO delays to show that the implied underlap design tradeoff for ultimate CMOS speed is affected significantly by parasitic G-S/D capacitance, i.e., fringe capacitance, in nanoscale devices.

As mentioned, DG FinFETs with undoped UTBs are very attractive for scaled CMOS due to their inherent benefits, i.e., better SCE control, smaller intrinsic gate capacitance in weak/moderate inversion, and high $I_{on}/I_{off}$ ratio. However, with the ultimate limit of UTB, i.e., ~5nm [Tri03a] due to severe quantization effects and technological difficulties, DG FinFET scaling to and beyond the HP25 node with $L_g = 10$nm [Sem05] seems to be extremely difficult since the fin thickness required for SCE
control is \( w_{Si} \equiv L_{eff}/2 \) [Yan05] if high-k gate dielectric is not viable. Thus, for further gate length scaling to 10nm and beyond, nonclassical CMOS devices have to be designed with the G-S/D underlap [Tri05a]. Even for the \( L_g > 10 \text{nm} \) regime or/and when a reliable high-k gate dielectric is developed, the underlap structure should be quite useful in the device design for effecting an optimal SCEs vs. \( I_{on} \) trade-off [Kra06, Lim05, Tri05a].

This benefit of the underlap structure in the DG FinFET should be most useful for SRAM applications. This is because high \( V_t \) tends to give large read static noise margin (read-SNM) and write-margin [Guo05], and can be easily obtained by SCE control via the effective channel length \( (L_{eff}) \) modulation in the weak-inversion region [Fos03c]. On the other hand, for the device speed issue, with the insight gained from the relationship between the S/D doping profile and \( V_t \) (and thus \( I_{on} \)), we can minimize the \( I_{on} \) loss, still keeping the parasitic capacitance small by controlling the extension length \( (L_{ext}) \) and straggle \( (\sigma_L) \). Thus, the underlap can also be quite useful in improving the device speed.

In chapter 5, we first explore SRAM cell design and scaling via DG FinFETs with G-S/D underlap. For this study, DG FinFETs with underlap are first characterized in terms of \( V_t \) for various \( L_{ext}, \sigma_L, \) and \( w_{Si} \) via 2-D numerical [Med04] and analytical simulations [Fos06a]. The relationship between \( V_t \) and read-SNM is verified to define an optimal SRAM cell, for the HP45 node with \( L_g = 18 \text{nm} \) [Sem05], with large read-
SNM and write-margin as well as less sensitivity to process variations of \( L_{\text{ext}} \) and \( \sigma_L \). Then, a scalability study of DG FinFET-based SRAM cell, with and without the G-S/D underlap, is done. Finally, based on the insight gained from \( V_t \) shift and \( I_{\text{on}} \) variation by \( \sigma_L \) changes, we optimally design DG FinFETs to improve the device speed.

In Chapter 6, this dissertation is concluded with a summary and suggestions for future works.

Appendixes A and B describe supporting UFDG model studies and a unique DG MOSFET feature, respectively.
CHAPTER 2
NANOSCALE CMOS: POTENTIAL NONCLASSICAL TECHNOLOGIES
VERSUS A HYPOTHETICAL BULK-SILICON TECHNOLOGY

2.1 Introduction

For classical bulk-Si and partially depleted (PD) SOI CMOS, continued scaling much beyond a physical gate length of \( \sim 50 \text{nm} \) [Sem01] is doubtful because of severe short-channel effects (SCEs) and unacceptably low \( I_{\text{on}}/I_{\text{off}} \) ratio. Hence, there is a growing interest in nonclassical fully depleted (FD) SOI single-gate (SG) and double-gate (DG) MOSFETs with ultra-thin bodies (UTBs) because of their inherent suppression of SCEs. Further, their small intrinsic gate capacitance in weak/moderate inversion and, especially for DG devices, the high \( I_{\text{on}}/I_{\text{off}} \) ratio stemming from the nearly ideal subthreshold gate swing imply substantial CMOS speed superiority over the classical SG counterparts [Fos02]. However, DG technology is complex; the DG FinFET [His98, Hua99] is easiest to fabricate, but its proven utility is years away. Contrarily, FD/SOI SG technology is less complicated; UTBs and metal gates are the main obstacles in its development [Cho00].

Because of the technological complexities and difficulties associated with DG CMOS, questions have been posed about the performance advantage relative to SG CMOS that it can potentially provide. For example, if the DG MOSFET gives twice the current, but with
twice the gate capacitance, then excessive device parasitics implied by the complex technology might render inferior performance. Further, it has been argued that SCEs in the bulk-Si SG MOSFET could be effectively suppressed by super-halo channel doping such that bulk-Si CMOS could actually be scaled down to 25nm channel lengths [Tau98]. However, this argument is simulation-based, and there is uncertainty about the physical modeling assumed [Ge01] and whether the assumed device structure could even be fabricated [Tau98]. Nonetheless, given such a “hypothetical” nanoscale bulk-Si CMOS technology, more detailed insights on the relative performance potentials of nonclassical UTB CMOS would be useful in deciding how and if they should be aggressively pursued.

In this chapter, we first compare asymmetrical and symmetrical DG, and FD/SOI SG devices via simulations done with our process/physics-based compact model UFDG [Fos03a, Fos04a], with emphasis on their immunity to SCEs. Then, using UFDG, we optimally design these nonclassical devices for 28nm gate length, and project their characteristics, including CMOS ring-oscillator delays, which we compare with projections of a hypothetical 28nm bulk-Si SG CMOS design derived, using our UFPDB compact model [Fos97], from Taur et al. [Tau98]. Interpretations of the simulation results give good physical insights on the nonclassical technologies, and indicate which ones might best replace the classical technologies at nanoscale nodes of the SIA ITRS [Sem01].
2.2 UFDG, UFPDB, and Simulation Conditions

The UFDG model [Fos03a, Fos04a] is generic, and hence applicable to FD/SG MOSFETs with relatively thick or thin [Fen03] back gate oxide as well as to both asymmetrical (ADG) and symmetrical (SDG) DG MOSFETs. The process/physics basis of UFDG makes it predictive and useful for doing the comparative projections. The model parameters can be defined largely from the device structure and physics. UFDG predicts SCEs mainly via a 2D solution of Poisson’s equation in the UTB for weak-inversion conditions [Yeh95]. In the strong-inversion region, the quantum-mechanical carrier confinement is incorporated in UFDG via the derivation of an iterative, self-consistent solution, dependent on both gate voltages \( V_{Gfs} \) and \( V_{Gbs} \), of the 1D (in x) Schrödinger and Poisson equations in the UTB/channel [Ge02]. This solution further physically accounts for the charge coupling between the front and back gates, and properly models the charge distribution throughout the Si-film UTB. UFDG also accounts for the dependences of carrier mobility on the UTB thickness \( t_{Si} \) as well as on the transverse electric field \( E_x \), and quasi-ballistic carrier transport in scaled devices is modeled via carrier-velocity overshoot, which is characterized in terms of carrier temperature [Ge01]. In addition, UFDG includes the parasitic (coupled) BJT (current and charge) and temperature dependence without the need for any additional parameters.

The UFPDB model [Fos97] is also process/physics-based, and is unified for application to PD/SOI as well as bulk-Si SG MOSFETs. This
model is based on a presumed retrograded, or super-haloed, channel, and it physically accounts for polysilicon-gate depletion [Chi01], carrier-energy quantization [Chi01], carrier velocity saturation with possible overshoot [Ge01], and gate-body tunneling current [Yan04].

For the nonclassical devices, we assume undoped bodies since, technologically, dopant control in UTBs is virtually impossible, as in extremely scaled bulk-Si and PD/SOI MOSFETs. Then, since the number of natural dopants in the body will actually be zero when device dimensions are extremely scaled, the UTBs are modeled as intrinsic. However, to avoid numerical instabilities in UFDG, \( N_B = 10^{15}/\text{cm}^3 \) is used for the simulations. Indeed, such a small \( N_B \) is virtually equivalent to \( N_B = 0 \) [Tri03a]. For SDG and FD/SG devices, we select metal gates (for \( V_t \) control), while \( n^+ \)- and \( p^+ \)-polysilicon gates are assumed for ADG devices, but without accounting for gate-depletion effects (which is justified somewhat by the relatively low \( E_x \)). The front and back gate-oxide thicknesses (or EOTs) are assumed equal in the DG devices, except for a brief analysis of the effects of unequal thicknesses in ADG devices. For the FD/SG devices, a thick buried oxide (\( t_{\text{BOX}} = 200\text{nm} \)) on a lightly doped (\( 10^{15}/\text{cm}^3 \)) p-type Si substrate is assumed. We ignore the source/drain field fringing in the BOX [Yeh95], which can exacerbate the SCEs that are predominantly governed by the UTB. Hence, it should be noted that the actual effective channel lengths (\( L_{\text{eff}} \)) of given FD/SG devices might be a bit longer (by \( \sim 3-5\text{nm} \)) than the values stated. However, if the actual
devices were designed with gate-source/drain underlap [Tri05a], which will probably be necessary, then the results of our study are reasonable as given. The effects of thinning $t_{\text{BOX}}$ are subsequently examined. For the bulk-Si/SG CMOS, $n^+$- and $p^+$-polysilicon gates, with gate-depletion effects, are assumed for nMOSFET and pMOSFET, respectively. We employ a retrograded channel, which adequately reflects the super-halo doping suggested by Taur et al. [Tau98], and $V_t$ is tuned via the lower, surface doping density ($N_{BL}$ in UFPDB).

2.3 Immunity to Short-channel Effects

We first compare SCEs in the nonclassical DG and FD/SG MOSFETs. Using UFDG, we vary the gate length ($L_g$, assumed to equal $L_{\text{eff}}$) from 500nm to 28nm, fixing $t_{\text{Si}}$ and the gate-oxide thickness ($t_{\text{oxf}}$ and $t_{\text{oxb}}$ for the DG devices) at 10nm and 3nm, respectively. The predicted threshold voltage roll-offs ($\Delta V_t(L_g)$ at $V_{DS}=50\text{mV}$) of the ADG, SDG, and FD/SG n-channel devices are compared in Fig. 2.1. Here, for each device, $V_t$ is defined via $I_{DS}(V_{GS}=V_t) = 10^{-7}W_g/L_g \text{ (A)}$, and $V_t$ of the $L_g = 500\text{nm}$ device is taken as the reference. As indicated in the figure, for the ADG and SDG devices, $\Delta V_t$ is negligible and no significant differences between the two DG devices are evident down to $L_g \sim 70\text{nm}$. However, when $L_g$ is scaled below $\sim 70\text{nm}$, the ADG MOSFETs show superior $\Delta V_t$ control over the SDG counterparts. This can be explained by the higher $E_x$ and stronger gate-gate charge coupling in the ADG devices [Kim01]. For the FD/SG MOSFETs, as clearly shown in Fig. 2.1, the $V_t$ roll-off is much more
Figure 2.1 UFDG-predicted threshold voltage roll-off (solid line) and DIBL (dash line) versus gate length (=L_{eff}) of the ADG, SDG, and FD/SG n-channel devices with t_{oxf}=t_{oxb}=3nm (for FD/SG, t_{BOX}=200nm) and t_{Si}=10nm; V_{t} is defined via I_{DS}(V_{GS}=V_{t}) =10^{-7}W_{g}/L_{g} [A] at V_{DS}=50mV, and, for each device, V_{t} of the L_{g}=500nm device is taken as the reference. DIBL is defined via \Delta V = V_{t}(V_{DS}=50mV)-V_{t}(V_{DS}=1.0V).
severe than those of the DG counterparts, implying that much thinner UTB will be necessary to control the SCEs in FD/SG CMOS.

Also in Fig. 2.1, predicted DIBL (defined by $\Delta V = V_t(V_{DS}=50\text{mV}) - V_t(V_{DS}=1.0\text{V})$) is given, showing similar superiority of the DG devices over the FD/SG device. And, as for $\Delta V_t$, the ADG device is better than the SDG device with regard to $\Delta V$. The predicted subthreshold gate swing ($S$) shows trends similar to those of $\Delta V_t$ and $\Delta V$ shown in Fig. 2.1. Also, we observe that for long $L_g (> \sim 70\text{nm})$, $S$ approaches its ideal value ($60\text{mV}$ at $300\text{K}$) for both DG devices, and nearly so ($61\text{mV}$) for the FD/SG device since the body-effect coefficient ($m = 1 + C_{\text{Body}}/C_{\text{oxf}} = 1 + C_{\text{oxb}}/C_{\text{oxf}} [\text{Lim85}]$) is only slightly greater than unity due to the thick $t_{\text{BOX}}$.

### 2.4 28nm Device Design

To compare classical and nonclassical CMOS performance potentials, we first use UFDG and UFPDB to optimally design the devices at the $L_g = 28\text{nm}$ node. Consider the ADG nMOSFET initially. To reduce the SCEs reflected by Fig. 2.1 to acceptable values, two approaches are checked: thinning the gate oxide and thinning the Si-film thickness. UFDG shows that when $t_{\text{oxf}} = t_{\text{oxb}} = t_{\text{ox}}$ is decreased from $3\text{nm}$ to $1\text{nm}$ in the ADG nMOSFET ($t_{\text{Si}}$ remains fixed at $10\text{nm}$), the SCEs ($\Delta V$ and $S$) are steadily and significantly improved. Further, because of the lowering of $S$, $I_{\text{on}}$ is continually increased, and $I_{\text{off}}$ is decreased. However, thinning the oxide thickness toward $1\text{nm}$ is not necessarily optimal due to the abrupt increase of gate-tunneling leakage current [Gha00, Yan04]. Indeed, for the
ADG device, limiting $t_{ox}$ at 2nm might be, with thin $t_{Si}$, enough to suppress severe SCEs: we predict $\Delta V \sim 56$mV and $S \sim 76$mV for this $t_{ox}$.

However, with $n^+$- and $p^+$-polysilicon gates, this ADG device design, i.e., $t_{ox}=2$nm and $t_{Si}=10$nm, is not viable for high-performance (HP) CMOS applications [Sem01], even with modification of $t_{Si}$; the low $I_{off} (~5.0$nA/$\mu$m) reflects a too-low $I_{on}$. Rather, this design can be made applicable to low-operation-power (LOP) CMOS ($I_{off} \sim 0.8$nA/$\mu$m from the ITRS roadmap [Sem01]) by adjusting $t_{Si}$ to 8.6nm. Therefore, our pragmatic optimal ADG design is initially taken as $t_{ox} = 2$nm and $t_{Si} = 8.6$nm for LOP applications.

Analogous to thinning $t_{ox}$, decreasing $t_{Si}$ from 10nm to 6nm in the ADG device (with $t_{ox}$ held at 3nm) also yields steady improvement in SCE control (reductions in $\Delta V$ and $S$), but since the reduction in $S$ is relatively small, the increase in $I_{on}$ is not as significant as that obtained by thinning $t_{ox}$. We note further that $V_t$ and $I_{off}$ of the ADG device might be controlled, without much change in $I_{on}$, by adjusting $t_{Si}$, keeping the pragmatic $n^+$- and $p^+$-polysilicon gates, instead of varying the channel doping, which is not viable.

To assess the two noted approaches for optimizing the ADG device design, we show in Fig. 2.2 UFDG-predicted $I_{DS}(V_{GS})$ characteristics of the initially optimized device and of one designed with the same $I_{off}$ via thinner $t_{Si}$ (7.4nm) and thicker $t_{ox}$ (3nm). As evident in this figure, the former device, with the thinner $t_{ox}$, shows lower $S$ and
Figure 2.2 UFDG-predicted current-gate voltage characteristics, at low and high drain voltages, of the initially optimized 28nm ADG device and of one designed with the same I_{off} (~0.76nA/mm) via thinner t_{Si} and thicker t_{ox}.
higher $I_{on}$, and hence it is solidified as our optimal 28nm ADG MOSFET for LOP applications.

To gain more insight on this ADG device design, we show in Fig. 2.3 predicted effects of separately varying $t_{oxf}$ and $t_{oxb}$. In Fig. 2.3(a), we see that varying only $t_{oxf}$ offers some control of $V_t$. However, we also see that $S$ and $I_{on}$ are degraded when using thicker $t_{oxf}$ due to the decreased sensitivity of the body potential to the voltage change on the front gate. Interestingly, however, in Fig. 2.3(b) we see that varying only $t_{oxb}$ is more effective in controlling $V_t$, with smaller effects on $S$ and $I_{on}$. This is because the ADG MOSFET has only one predominant channel, which is closer to the front (n$^+$) gate [Kim01]. Thus, for ADG devices with n$^+$- and p$^+$-polysilicon gates, varying $t_{oxb}$, with fixed $t_{oxf}$ for SCE control, might be useful for $V_t$ tuning. However, for low-standby-power (LSTP) applications with very low $I_{off}$ $\sim$ 1pA/µm [Sem01], this approach to $V_t$ control should also include $t_{Si}$ variation, as indicated in Fig. 2.3(b).

Now, for a fair comparison, we design the SDG (e.g., a FinFET) and FD/SG nMOSFETs with the same structure as the optimized ADG device, i.e., with $t_{ox}=2$nm and $t_{Si}=8.6$nm, and the same $I_{off}$ ($\sim$0.8nA/µm) for LOP applications. We get the specified $I_{off}$ by tuning the metal-gate work functions (which might not be so easily done technologically): $\Phi_M = 4.62$V for the SDG device, and $\Phi_M = 5.02$V for the FD/SG device. And, knowing the FD/SG MOSFET will show inferior SCE control, we also define a more aggressive, optimal version of it with $t_{ox} = 1$nm and $t_{Si} = 5$nm, which are
Figure 2.3  UFDG-predicted effects of (a) separately varying $t_{\text{oxf}}$ in the ADG device with $t_{\text{oxb}}$ fixed at 2nm and (b) separately varying $t_{\text{oxb}}$ in the ADG device with $t_{\text{oxf}}$ fixed at 2nm, showing some control of both $V_t$ and $S$, and how the effects are enhanced when $t_{\text{Si}}$ is thinned to 6nm (shown in (b)).
close to the technological limits of these structural parameters. This optimal device is also designed, neglecting the possible significance of gate-tunneling leakage, to have the same $I_{\text{off}}$ (with $\Phi_M = 4.51\text{V}$) as that of the optimized ADG device.

For the hypothetical bulk-Si/SG nMOSFET design, we use UFPDB, generally following Taur et al. [Tau98] for SCE control, but using a more aggressively scaled gate oxide, $t_{\text{ox}} = 1\text{nm}$ as for the optimal FD/SG device, with gate tunneling still neglected. The retrograded channel is defined with surface doping density $N_{\text{BL}} = 3.42 \times 10^{18}/\text{cm}^3$, a peak body doping density ($N_{\text{BH}}$) of $10^{19}/\text{cm}^3$, and an effective depletion thickness ($T_B$) of 14.2nm. The noted $N_{\text{BL}}$ of this nMOSFET, with $n^+$-polysilicon gate, was tuned to yield $I_{\text{off}}$ equal to that of the nonclassical devices. Note that this channel/body doping profile is probably not manufacturable, rendering a hypothetical device.

Figure 2.4 shows UFDG/UFPDB-predicted $I_{\text{DS}}(V_{\text{GS}})$ characteristics of the two optimal DG devices, of the two designed FD/SG devices, and of the bulk-Si/SG device. Clearly, the suboptimal FD device with the thicker $t_{\text{Si}} (= 8.6\text{nm})$ suffers from severe SCEs; $\Delta V \equiv 266\text{mV}$ and $S \equiv 108\text{mV}$. However, the SDG device, like the ADG device with the same $t_{\text{Si}}$, shows good control of the SCEs, with the ADG device being a bit superior in this regard; $\Delta V$ is 42mV and 72mV for the ADG and SDG devices, respectively, and $S$ is 72mV and 74mV. For the optimal, thinner-$t_{\text{Si}} (= 5\text{nm})$ FD/SG device, good SCEs ($\Delta V \equiv 43\text{mV}$ and $S \equiv 71\text{mV}$) and larger
Figure 2.4 UFDG/UFPDB-predicted current-gate voltage characteristics of the two optimally designed DG nMOSFETs, with $t_{ox}=2\text{nm}$ and $t_{Si}=8.6\text{nm}$, of the two designed FD/SG nMOSFETs, i.e., optimal FD/SG with $t_{oxf}=1\text{nm}$, $t_{BOX}=200\text{nm}$, and $t_{Si}=5\text{nm}$, and suboptimal FD/SG with $t_{oxf}=2\text{nm}$, $t_{BOX}=200\text{nm}$, and $t_{Si}=8.6\text{nm}$, and of the optimally designed bulk-Si/SG nMOSFET, with $t_{ox}=1\text{nm}$ and $T_{B}=14.2\text{nm}$, at high drain voltage; for all devices, $L_g=28\text{nm}$ and $I_{off}=0.76\text{nA}/\mu\text{m}$.
Ion, relative to the suboptimal design, are predicted. We stress, however, that if the DG devices were thinned to $t_{Si} = 5$nm, their SCEs would be virtually nonexistent (as UFDG simulations show). For the bulk-Si/SG device, the SCEs are effectively suppressed due to the thin $T_B$ as shown in the Fig. 2.4. Note here that since DIBL renders a minor effect on the delay performance [Tau98], we selected $T_B$ to get relatively small $S$ ($\sim 80$mV), sacrificing DIBL ($\sim 120$mV) and necessitating the high $N_{BL}$ to keep $I_{off}$ under control.

As mentioned, the applications of the designed ADG device, with $n^+$- and $p^+$-polysilicon gates, are limited to LOP. Thus, for HP applications, we can consider only the defined SDG and FD/SG devices with new metal gates ($\Phi_M$ is 4.39V for the SDG device, and 4.69V and 4.29V for the suboptimal and optimal FD/SG devices, respectively), and the hypothetical bulk-Si/SG device with new $T_B$ (17.5nm, to keep $S \sim 80$mV) and $N_{BL}$ ($10^{18}$/cm$^3$) for appropriate $I_{on}/I_{off}$.

For the CMOS technologies, the SDG and FD/SG pMOSFETs are designed to have the same $I_{off}$ as that of the ADG nMOSFET by using metal gates and tuning the work functions. However, for the ADG pMOSFET, $n^+$- and $p^+$-polysilicon gates are still assumed, but switched for the back and front gates, respectively. Thus there is a slight discrepancy in $I_{off}$ between the two ADG CMOS devices because of the different electron and hole mobilities. The bulk-Si pMOSFETs are also designed to have the specified $I_{off}$ by adjusting $N_{BL}$ with a $p^+$-polysilicon gate, keeping
TB and NBH the same in the nMOSFETs. For ring-oscillator simulations, we assume the source/drain areas of the bulk-Si CMOS devices are defined based on lengths of $3 \times \text{(pitch/2)}$ [Sem01].

2.5 CMOS Performance Projections

To compare CMOS speeds, 9-stage unloaded CMOS-inverter ring oscillators (ROs) were simulated with UFDG and UFPDB in Spice3. Predicted propagation delays for the classical and the four nonclassical device designs (for LOP) are plotted in Fig. 2.5 versus supply voltage $V_{DD}$. As expected, the DG CMOS designs are faster than the SG ones, including bulk-Si CMOS, over the entire voltage range, while the ADG and SDG CMOS delays are virtually the same. Interestingly, the optimal FD/SG design has comparable speed to the DG CMOS designs: only $\sim$15% longer delay at $V_{DD} = 1.2V$. Further, the speed of the suboptimal FD/SG design is not much worse at high $V_{DD}$ ($\sim$34% slower at $V_{DD} = 1.2V$, compared to the DG CMOS), although the delays are significantly longer at $V_{DD} < \sim$1.0V. Contrarily, the predicted RO delay for the bulk-Si SG CMOS is much longer than those of all the nonclassical CMOS designs over the entire voltage range, even though the bulk-Si devices are hypothetical and seemingly optimal.

For HP-applicable CMOS, i.e., the SDG, the FD/SG, and the bulk-Si/SG designs all with $I_{off} = 0.7 \mu A/\mu m$ as noted, the RO simulation results, shown in Fig. 2.6, are very interesting. As in the LOP CMOS RO results, the SDG design shows speed superiority over the optimal FD/SG
Figure 2.5 UFDG/ and UFPDB/Spice3-predicted propagation delays versus supply voltage of 9-stage unloaded CMOS-inverter ring oscillators comprising the five 28nm DG, FD/SG, and bulk-Si/SG LOP device designs. Gate-source/drain overlap of 10% of $L_g$ was assumed for all gates. Three of the five delay curves are re-plotted in the inset for better view. The off-state current of all devices was matched to 0.8nA/mm.
Figure 2.6  UFDG/ and UFPDB/Spice3-predicted propagation delays versus supply voltage of 9-stage unloaded CMOS-inverter ring oscillators comprising the four 28nm SDG, FD/SG, and bulk-Si/SG HP device designs. Gate-source/drain overlap of 10% of $L_g$ was assumed for all gates. The off-state current of all devices was matched to 0.7mA/mm.
and the bulk-Si/SG designs over the entire voltage range. Surprisingly, however, the suboptimal $t_{Si} = 8.6\text{nm}$ FD/SG design actually yields shorter delay relative to that of the SDG design for $V_{DD} > \sim 0.95\text{V}$; but for lower $V_{DD}$, it becomes slower. The equivalent speed performance for higher $V_{DD}$ seems inconsistent with the predicted currents of the FD/SG and SDG devices for HP, which can be inferred by shifting the $I_{DS}(V_{GS})$ characteristics for LOP in Fig. 2.4. The FD/SG device shows much lower $I_{on}$ (by $\sim 37\%$ at $V_{DD} = 1.0\text{V}$).

To explain these unexpected suboptimal FD/SG RO results, we consider intrinsic gate capacitance of the nonclassical devices. In the subthreshold region, the inversion charge is negligible, and thus the gate capacitance of the intrinsic FD/SG MOSFET can be expressed by the series combination of oxide capacitance and the effective body capacitance [Lim85]:

$$C_G = \left( \frac{1}{C_{oxf}} + \frac{1}{C_{Body}} \right)^{-1} \equiv C_{Body} \quad (2.1)$$

where $C_{oxf}$ is the front gate oxide capacitance and $C_{Body}$ includes the UTB depletion capacitance, $C_b = \varepsilon_{Si}/t_{Si}$, and the buried oxide capacitance, $C_{oxb} = \varepsilon_{ox}/t_{oxb}$:

$$C_{Body} = \frac{C_{oxb}C_b}{C_{oxb} + C_b} \equiv C_{oxb} \ll C_b, C_{oxf} \quad (2.2)$$

For the common thick $t_{oxb}$ for FD/SG, i.e. $t_{BOX}$, $C_{Body}$ in (2.2) is relatively small, and hence so is $C_G$ in (2.1). Indeed, the subthreshold-region gate
capacitance of the FD/SG device is defined predominantly by parasitic capacitance, e.g., the gate-source/drain overlap capacitance. Further, for \( V_{GS} \) increasing to strong inversion, the increase in \( C_G \) defined by the inversion-charge is, in our suboptimal \( t_{Si} = 8.6 \text{nm} \) FD/SG device, deferred to higher voltages (\( V_{GS} \sim 0.75 \text{V} \)) because of the high \( S \). Therefore, the device shows extremely small \( C_G \) (including the parasitics) at low \( V_{GS} \) (\( \sim 0.5 \text{V} \)), as shown by the UFDG-predicted curve in Fig. 2.7. However, when we thin \( t_{Si} \) to get better control of the SCEs, \( S \) decreases, and the optimal FD/SG device shows, also in Fig. 2.7, intermediate-\( V_{GS} \) \( C_G \) that is much larger than that of the suboptimal FD/SG device. Similarly, the DG devices show in Fig. 2.7, because of device neutrality [Fos02], very small \( C_G \) at low \( V_{GS} \), but the increase due to inversion charge occurs at lower \( V_{GS} \) due to the low \( S \). The comparative result then is that the DG devices show much higher gate capacitance than the suboptimal FD/SG device at all \( V_{GS} \). We believe that the relative \( C_G(V_{GS}) \) curves in Fig. 2.7, irrespective of the \( I_{DS}(V_{GS}) \) characteristics inferred from Fig. 2.4, underlie the surprising RO results in Fig. 2.6.

Unlike the nonclassical devices, \( C_{Body} \) of the classical bulk-Si/SG device is defined by the large depletion capacitance (i.e., \( \varepsilon_{Si}/T_B \)), and hence \( C_G \) is finite and substantive as indicated by the UFPDB-predicted curve in Fig. 2.7. Because of the high \( C_G \) in the weak/moderate inversion region, in addition to the areal source/drain junction capacitance, and the polysilicon-gate depletion effect in strong inversion, the bulk-Si CMOS
Figure 2.7  UFDG- and UFPDB-predicted gate capacitances versus gate voltage, at low drain voltage, for the four 28nm SDG, FD/SG, and bulk-Si/SG device designs. Gate-source/drain overlap of 10% of L_g was assumed for all gates.
speed is much slower than those of the nonclassical CMOS designs. The main effect of the gate depletion is reduced $I_{on}$, as reflected by the decreasing $C_G$ for increasing $V_{GS}$ in Fig. 2.7.

The propagation delay reflected by the oscillation frequency of the RO is defined by the pull-down ($V_{DD}$-to-$V_{DD}/2$) and pull-up (0-to-$V_{DD}/2$) times of a constituent inverter. These times depend on the $V_{DS}$-dependent currents ($I_{DS}(t)$) in the driving transistors, and on the capacitive load at the output terminal, which, for the unloaded RO, is predominantly the sum of the nMOSFET and pMOSFET gate capacitances of the next stage ($C_{Gn}(t) + C_{Gp}(t)$). We define, for pull down in the SDG and the two FD/SG CMOS designs, the dynamic current $I_{DS}(t)$ of the driving transistor between $V_{DS}(t) = V_{DD}$, which corresponds to $V_{GS}(t) = V_{DD}/2$, and $V_{DS}(t) = V_{DD}/2$. Similarly, we define the dynamic charging current ($I_Q(t)$) at the inverter-output node, which is the gate current of the next-stage inverter. With these two dynamic currents $I_{DS}(t)$ and $I_Q(t)$, we can estimate the pull-down time ($t_{pd}$, which is comparable to the pull-up time $t_{pu}$) of a constituent inverter for each design, and thereby confirm our belief about the surprising relative delays in Fig. 2.6.

To simplify our estimations, we define an average value of the total gate-capacitance load,

$$
\bar{C}_G = \left( \frac{2}{V_{DD}} \right) \int_{0}^{t_{pd}} I_Q dt,
$$

(2.3)

and use it to approximate the pull-down delay as
Indeed, for \( V_{\text{DD}} = 0.65 \text{V} \), we get using (2.3) and (2.4) \( \tau_{\text{pd}} \cong 2.43 \text{ps} \) and 1.93ps for suboptimal FD/SG and SDG, respectively. However, for \( V_{\text{DD}} = 1.2 \text{V} \), we find \( \tau_{\text{pd}} \cong 1.49 \text{ps} \) and 1.43ps for FD/SG and SDG, respectively. (Note that the estimated delays are a bit shorter than the predicted RO delays \( (=t_{\text{pd}}+t_{\text{pu}})/2 \) in Fig. 2.6 due to the neglected parasitic capacitances such as the gate-source/drain overlap capacitances.) Thus, these estimated values of \( t_{\text{pd}} \) are in accord with the surprising results in Fig. 2.6, i.e., the significant SDG speed superiority at low voltages and the comparable FD/SG speed at high voltages.

Now, by defining an average \( \overline{I_{\text{DS}}} \) in (2.4) analogous to \( C_{\text{G}} \) in (2.3), we can evaluate the contributions of the driving current and the capacitive load in determining the RO delays for each CMOS design. For \( V_{\text{DD}} = 0.65 \text{V} \), we get the average-current ratio \( \overline{I_{\text{DS}}}^{(\text{SDG})}/\overline{I_{\text{DS}}}^{(\text{FD})} \cong (\overline{I}/I_{\text{DS}}^{(\text{FD})})/(\overline{I}/I_{\text{DS}}^{(\text{SDG})}) = 3.91 \), where FD here refers to the suboptimal \( t_{\text{Si}} = 8.6 \text{nm} \) design, and the average-capacitance ratio \( C_{\text{G}}^{(\text{SDG})}/C_{\text{G}}^{(\text{FD})} = 3.09 \). For \( V_{\text{DD}} = 1.2 \text{V} \), we get \( \overline{I_{\text{DS}}}^{(\text{SDG})}/\overline{I_{\text{DS}}}^{(\text{FD})} = 2.24 \) and \( C_{\text{G}}^{(\text{SDG})}/C_{\text{G}}^{(\text{FD})} = 2.20 \). These ratios, with reference to (2.4), explain that the SDG speed superiority at low voltages comes from the relatively high average drive current (i.e., \( \overline{I_{\text{DS}}}^{(\text{SDG})}/\overline{I_{\text{DS}}}^{(\text{FD})} > C_{\text{G}}^{(\text{SDG})}/C_{\text{G}}^{(\text{FD})} \)) in the SDG devices. Also, they explain that the surprising comparable FD speed at high \( V_{\text{DD}} \) is due to the relatively low average gate capacitance (i.e., \( \overline{I_{\text{DS}}}^{(\text{SDG})}/\overline{I_{\text{DS}}}^{(\text{FD})} \equiv \))
Other RO simulations done with UFDG/Spice3 reveal that this FD/SG speed merit at high $V_{DD}$ is maintained when the gate-source/drain overlap capacitances are increased up to 30%. However, for the optimal FD/SG design with $t_{Si} = 5$nm, we find for all $V_{DD}$ that $\overline{C_G}$ and $\overline{1/I_{DS}}$ are comparable to the corresponding averages of the DG device (i.e., $\overline{C_{G(SDG)}}/\overline{C_{G(FD)}}$ is 1.06 and 1.08 and $\overline{I_{DS(SDG)}}/\overline{I_{DS(FD)}}$ is 1.06 and 1.11 for low and high $V_{DD}$, respectively). As a result, we get, from the average $\overline{C_G}$ and $\overline{1/I_{DS}}$ (or directly using (2.4)), the comparable $t_{pd}$ for the optimal FD/SG compared to SDG. All the relative RO delays of the nonclassical CMOS in Fig. 2.6 are hence explained, and good physical insights regarding them is attained.

The speed comparisons in Figs. 2.5 and 2.6 were derived from unloaded RO simulations. With loading, we anticipate that the relative performance of the suboptimal FD/SG CMOS will deteriorate because of its lower current drive. The UFDG/Spice3-predicted loaded ($C_L$ on each stage) RO delays plotted in Fig. 2.8 versus $C_L$ confirm this anticipation. Hence, in general applications, DG CMOS should be substantively faster than the FD/SG counterpart, especially for heavy loads and low supply voltages. Nevertheless, it is interesting to learn that for light loads and HP applications, suboptimal FD/SG can yield speeds comparable to DG.
Figure 2.8 UFDG/Spice3-predicted loaded ($C_L$ on each stage) CMOS ring-oscillator delays versus $C_L$, at $V_{DD}=1.2V$, for the 28nm SDG and FD/SG device designs.
2.6 Thin-BOX FD/SOI CMOS

For the FD/SOI SG devices with thick BOX, it has been reported that the BOX field fringing and the coupling between the source/drain and channel will degrade the SCE control, and thus using thin BOX (< ~40nm) for nanoscale FD/SOI devices is beneficial for improving SCEs and the $I_{on}/I_{off}$ ratio [Fen03]. Also, it was confirmed [Num02] that whereas for long-$L_g$ devices $S$ increases with decreasing $t_{BOX}$, $S$ for shorter $L_g$ is minimized for $t_{BOX} < 50$nm. However, the benefit of thinning $t_{BOX}$ is much reduced when $t_{Si}$ is ultrathin [Tri03a] since the underlying fringing-field effect in the BOX is reduced along with the SCEs via the UTB. Also, when the BOX is thinned, the effective body capacitance [Lim85] and the source/drain junction capacitance [Yeh95] will be increased, implying that thin BOX may undermine FD/SOI CMOS speed. Note also that the increased $C_{Body}$ implies added sensitivity of the FD/SG device characteristics, e.g., $V_t$ [Lim85], to variations in $t_{Si}$.

To solidify our notion [Tri03a] that thinning the BOX is not a judicious design option, we examine its effects more closely using UFDG. When the BOX is thinned down to 20nm in our optimal FD/SG devices, UFDG predicts that the intrinsic gate capacitance $C_G$ is increased a bit, especially in the subthreshold region, while the average $\overline{C}_G$ from RO simulation stays almost constant for $V_{DD} = 1.2$V. The increase can be explained by the increment of $C_{Body}$ in (2.2) with reduced $t_{BOX}$ (or $t_{oxb}$). For $t_{BOX}$ thinned down even further, $C_G$ is estimated to be comparable to
that of the correspondingly scaled bulk-Si MOSFET, which is finite and substantive for low $V_{GS}$ [Fos02]. However, this estimation tends to be too high since UFDG does not account for substrate, or back-gate depletion under the (front) gate.

In addition to the $C_G$ increase, thinning the BOX also increases the (quasi-static) parasitic source/drain capacitances to values that are comparable to $C_G$ in the subthreshold region. This is because the substrate under the source/drain tends to be inverted in the nMOSFET and accumulated in the pMOSFET, yielding source/drain capacitances that are defined mainly by $C_{oxb}$. However, for high-speed transients such as the RO oscillations, the inversion charge under the source/drain of the nMOSFET cannot respond to the applied transient voltage, thus causing a deep-depletion condition in the substrate. Therefore, the parasitic source/drain capacitance can be neglected for the nMOSFET. The substrate accumulation charge in the pMOSFET can respond, however, and hence the parasitic source/drain capacitance in it is determined by $C_{oxb}$, which increases for decreasing $t_{BOX}$. Based on the noted increases in gate and source/drain capacitances, we can predict that the CMOS speed will be substantially degraded when the BOX is aggressively thinned as suggested in [Fen03]. The UFDG/Spice3-predicted RO delays plotted in Fig. 2.9 versus $t_{BOX}$ for the optimal FD/SG CMOS provide clear evidence of the speed degradation. We include in the figure predicted delays without accounting for the areal source/drain capacitance. For this case,
Figure 2.9 UFDG/Spice3-predicted propagation delays of 9-stage unloaded CMOS-inverter ring oscillators comprising the optimal 28nm FD/SG device design, with $t_{\text{BOX}}$ thinned down to 15nm, at various values of supply voltage. Predicted delays for no areal source/drain capacitance are also shown.
we see that there is negligible effect of varying $t_{\text{BOX}}$ on the RO delays. We thus conclude that the effect of the increased parasitic source/drain capacitance in the pMOSFET is predominant in defining the noted speed degradation caused by thinning the BOX.

2.7 Summary

Using our process/physics-based compact models (UFDG and UFPDB) in Spice3, we have projected device characteristics and CMOS performances of nonclassical UTB CMOS technologies (ADG, SDG, and two versions of FD/SOI, all of which will, generally, require metal gates with tuned work functions for $I_{\text{off}}$ control) optimized at the $L_g = 28\text{nm}$ node (where $L_{\text{eff}} = L_g$ was assumed), and compared them with that of classical, hypothetical bulk-Si CMOS at this node. Comparisons of predicted SCEs ($\Delta V_t$, $\Delta V$, and $S$) of nonclassical devices and speeds (RO delays) of the nonclassical and classical CMOS were made, and good physical insights regarding their relative characteristics were given. With the same UTB thickness, the DG devices were shown to be far superior to the FD/SG devices with regard to SCE control, and generally superior to SG devices, including bulk-Si devices, with regard to speed because of higher drive currents. However, an interesting insight was noted. For light loads and moderate supply voltages, a suboptimal FD/SG design (with the same $t_{\text{Si}}$) for both LOP and HP applications was found to yield speeds comparable to the DG designs, even though its current drives are much lower and its SCEs are much more severe. This surprising comparison was shown to be
a result of the FD/SG devices having much lower intrinsic gate capacitance, which is due to their thick BOX and higher subthreshold swing, and hence deferred onset of significant inversion-charge capacitance. At lower $V_{DD}$, however, the DG designs are much faster because of their much higher drive currents. When the FD/SG CMOS design was optimized by aggressive scaling of the UTB thickness, its high-$V_{DD}$ speed diminished (but was still comparable to that of DG CMOS) because of higher gate capacitance at intermediate gate voltages, while its low-$V_{DD}$ speed improved due to increased current. Compared to the nonclassical CMOS, the predicted delay of the bulk-Si/SG CMOS was much longer due to its high gate capacitance in the weak/moderate inversion region, in addition to the areal source/drain junction capacitance, and relatively low drive current limited by polysilicon-gate depletion. Finally, we used UFDG/Spice3 RO simulations to show that FD/SOI CMOS speed is degraded as the BOX is thinned, mainly because of increased source/drain capacitance in the pMOSFET, thereby suggesting that such thinning, aimed at improved control of field fringing in the BOX, is not a good design tradeoff.
CHAPTER 3
BULK INVERSION IN FINFETS AND IMPLIED INSIGHTS ON EFFECTIVE GATE WIDTH

3.1 Introduction

While the double-gate (DG) FinFET (Fig. 3.1(a)) has become a leading device option for future nanoscale CMOS, there is a technological limit to the aspect ratio ($R_f$) of the Si-fin height ($h_{Si}$) to the width ($w_{Si}$). Since $w_{Si}$ must be ultra-thin for good control of short-channel effects (SCEs) [Fos04b], this limit implies small effective gate width (commonly assumed to be $W_{eff} \approx 2h_{Si}$) and, ostensibly, low on-state current ($I_{on}$) per pitch. There is therefore interest in making the FinFET a triple-gate (TG) transistor by activating the top gate (Fig. 3.1(b)), yielding, from a surface inversion-charge perspective, $W_{eff} \approx 2h_{Si} + w_{Si}$ as is commonly assumed, and alleviating, with a doped fin-body, the thin-$w_{Si}$ requirement [Doy03]. However, because of fin-corner effects [Fos03b] and technological limitations [Tri03a], the fin-body must be left undoped, and so relaxation of the thin-$w_{Si}$ (i.e., UTB) requirement for SCE control for the TG FinFET, relative to that for the DG device, is minimal [Fos04b]. Nonetheless, the larger $W_{eff}$ for the TG device could mean significantly higher $I_{on}$ relative to that of the DG FinFET even when $w_{Si}$ is thin for good SCE control.

In this chapter, we use three-dimensional (3-D) numerical device simulations to examine the increase in $I_{on}$ ($\Delta I_{on}$) of TG FinFETs with
varying $R_f = h_{Si}/w_{Si}$ that results from activation of the top gate. From the surface inversion-charge perspective, we expect a relative increase in $I_{on}$ of $\Delta W_{eff}/W_{eff(DG)} = w_{Si}/(2h_{Si}) = 1/(2R_f)$. However, our simulation results contradict this expectation, and give interesting insights concerning fin-body bulk inversion in undoped FinFETs, even in the on-condition, and imply the consequent inappropriateness of the commonly defined $W_{eff}$ as an indicator of $I_{on}$. Based on proper (re)definition of effective gate width reflecting bulk inversion, we further examine the gate layout-area issue [Yan05] of FinFET CMOS.

### 3.2 Numerical Simulations

We first use Davinci [Dav03], a 3-D numerical device simulator, to simulate DG and TG n-channel FinFETs as illustrated in Fig. 3.1. We assume abrupt source/drain junctions, and a metallurgical, or effective, channel length ($L_{eff} = L_{gate}$) of 25nm. The gate-oxide thickness ($t_{ox} = EOT$) is 1.2nm and the buried-oxide (BOX) thickness is 200nm. For the DG devices, the top-gate oxide thickness is 50nm, which effectively negates the top gate electrode; it is $t_{ox}$ for the TG devices. Based on [Fos03b, Fos04b, Tri03a, Yan05], we assume undoped Si-fin bodies with $w_{Si} = 13nm$ ($=L_{eff}/2$) and various values of $h_{Si}$. For threshold-voltage ($V_T$) control, a midgap metal gate is assumed.

(We note that the carrier-transport modeling in Davinci is deficient for nanoscale FinFETs since it is based mainly on studies of single-gate (SG) bulk MOSFETs. For example, carrier mobility in UTB-fin
Figure 3.1 Two-dimensional cross-sectional view of (a) the DG FinFET, specifying x and z directions in the fin-body, (b) the TG FinFET, showing the fin-body dimensions, and (c) the DG FinFET without the top gate stack (thick oxide and metal electrode).
channels depends on $w_{Si}$ as well as the transverse electric field [Ess03, Tri04] as does the carrier-energy quantization [Ge02], and Davinci misses these $w_{Si}$ dependences. We hence ignore the quantization and use the standard, universal mobility model in Davinci, but stress that the relative results presented here are nonetheless meaningful.)

3.2.1 I-V Characteristics of DG and TG FinFETs

Davinci-predicted current-voltage characteristics of the DG and TG FinFETs with $h_{Si} = 39$nm (i.e., $R_f = 3$) are shown in Fig. 3.2. These characteristics show a relative increase in $I_{on}$ (at $V_{GS} = V_{DS} = 1.0$V) of only 5.4% in the TG device, much less than the expected 16.7% ($= 1/(2R_f)$). The inset of Fig. 3.2 shows semi-log plots of the current-voltage curves, revealing the subthreshold characteristics of the two devices. The TG-FinFET $V_t$ is only $\sim$10mV higher than that of the DG FinFET (corresponding to $\sim$15%-lower $I_{off}$). The small difference between the subthreshold characteristics does not explain the noted discrepancy in the relative $\Delta I_{on}$. Such a significant discrepancy is also predicted for other values of $R_f$, ranging from about 1 to 5 (non-integers because of finite mesh spacing for the numerical simulations), as illustrated in Fig. 3.3. Note for $R_f = 1$, the Davinci-predicted relative increase in $I_{on}$ due to the top gate is only 14.0%, as opposed to the expected 54.2%. That is, $I_{on}$ of the DG FinFET for this extreme case is almost 90% of that in the TG counterpart, which is consistent with numerical results of Burenkov et al. [Bur02]. Our
Figure 3.2 Davinci-predicted current-voltage characteristics of undoped n-channel DG FinFETs, with and without the top gate stack, and of the TG counterpart, all with $h_{Si} = 39\text{nm}$, $w_{Si} = 13\text{nm}$, $t_{ox} = 1.2\text{nm}$, $t_{BOX} = 200\text{nm}$, $L_{eff} = 25\text{nm}$, and midgap metal gate. The semi-log replots of the three curves in the inset show the subthreshold characteristics, and small variations in $I_{off}$ and $V_t$ among the three devices.
Figure 3.3 Davinci-predicted on-state ($V_{GS} = V_{DS} = 1.0\text{V}$) current increase ($\Delta I_{on}$) due to the top gate of the TG nFinFET, relative to the DG nFinFET current ($I_{on(DG)}$), versus the fin aspect ratio; $L_{eff} = 25\text{nm}$ and $w_{Si} = 13\text{nm}$. Also plotted is the $W_{eff}$-based expectation for the relative current increase, defined by $w_{Si}/(2h_{Si}) = 1/(2R_f)$. 
results here clearly show that the commonly defined $W_{\text{eff}}$ is not a valid indicator of relative values of $I_{\text{on}}$ in TG and DG FinFETs.

3.2.2 Electric-Field Fringing Effects

A possible explanation for these surprising results is that electric-field fringing from the sidewall gates above the fin in the DG FinFET (see Fig. 3.1(a)) induces significant inversion charge in the top fin surface. Indeed, exploitation of such field fringing has been proposed to effect a bottom gate extension [Par02]. To check this explanation, we simulated the DG FinFET with its top gate stack (thick oxide and metal electrode) removed as illustrated in Fig. 3.1(c). The predicted current-voltage characteristic, for $R_f = 3$, is included in Fig. 3.2. We see that the field-fringing effect is negligibly small and does not explain the relatively small increase in $I_{\text{on}}$ of TG FinFETs; $I_{\text{on}}$ of the complete DG FinFET (Fig. 3.1(a)) is only 1.5% higher than that of the same device without the top gate stack.

3.2.3 Bulk Inversion

Insight into the actual explanation for the interesting results in Figs. 3.2 and 3.3 is provided by the Davinci-predicted electron densities ($n$) in the three devices of Fig. 3.2 ($R_f = 3$). Based on the x-z coordinate system shown in Fig. 3.1(a), $n(x,z)$ at $V_{\text{GS}} = V_{\text{DS}} = 1.0\text{V}$, taken from the center of the channel ($y = L_{\text{eff}}/2$), is shown in Figs. 3.4 and 3.5. (These are classical solutions; the effects of quantization are noted later. Note that our use of the predicted carrier densities to give insight on the predicted
Figure 3.4 Davinci-predicted on-state electron density along the top fin surface, at the center of the channel \((y = L_{\text{eff}}/2)\), in the DG and TG nFinFETs of Fig. 3.2.
Figure 3.5  Davinci-predicted on-state electron density down the middle of the fin, at the center of the channel ($y = L_{eff}/2$), in the DG and TG nFinFETs of Fig. 3.2.
currents in the DG and TG devices further justifies our lack of concern about the deficiencies of the carrier-transport modeling in Davinci.) In Fig. 3.4, the variation of $n$ across the top surface ($n(x=0,z)$) in the DG FinFET without the top gate stack shows substantive inversion charge away from the sidewalls, i.e., volume, or bulk inversion. The predicted $n(x=0,z)$ of the complete DG FinFET shows a moderate increase due to the noted field fringing. (The top metal electrode does nothing, as we confirmed via simulation.) The full effect of the field fringing is reflected in Fig. 3.5 where the predicted electron density down the middle of the fin ($n(x,z=6.5\text{nm})$) is shown. The integrated inversion charge in both DG FinFET structures, however, reflects the small 1.5% benefit of the field fringing to $I_{on}$. For the TG FinFET in Figs. 3.4 and 3.5, we see higher electron density near the top fin surface due to the third top gate. But, as discussed with reference to Figs. 3.2 and 3.3, that benefit in the TG device is much less than that implied by the increased $W_{eff}$.

As indicated in Figs. 3.4 and 3.5 then, we infer that the results in Fig. 3.3, i.e., lower than expected $I_{on}$ in TG FinFETs relative to that in the DG counterparts, are due to the strong bulk inversion that occurs in the on-state condition. Note the high $n$ ($>2\times10^{18}\text{cm}^{-3}$) throughout the fin bulk, away from the surfaces, in all three device structures. The bulk-inversion charge in the DG FinFET contributes significantly to $I_{on}$, perhaps due in part to the fact that the electron mobility in the fin bulk.
(\(\mu_b\)) can be higher than that at the surfaces (\(\mu_s\)) [Ess03, Tri04] and hence the activation of the top gate is not very beneficial.

To give more quantitative explanation, we express the DG FinFET on-state current, separating out surface (\(Q_{is}\)) and bulk (\(Q_{ib}\)) components of inversion-charge density:

\[
I_{on(DG)} = W_{eff} Q_{is} v_s + h_{Si} Q_{ib} v_b
\]

(3.1)

where \(v_s\) and \(v_b\) represent the average carrier velocities at the fin surfaces and in the fin bulk, respectively. Note that the velocities depend on, in addition to \(\mu_s\) and \(\mu_b\), \(V_{DS}\) which controls the electric field \(E_y(x)\) and governs velocity saturation/overshoot along the channel. Actually, (3.1) is a reasonable expression if \(R_f\) is greater than one, which makes the effective width of the fin-bulk component approximately \(h_{Si}\). For the \(R_f = 3\) DG FinFET of Figs. 3.4 and 3.5 (with \(w_{Si} = 13\)nm), we find surprisingly large \(Q_{ib} > Q_{is}\) at \(V_{DS} = V_{GS} = 1.0\)V, which, via (3.1), defines a predominant enhancement of \(I_{on(DG)}\) over that implied by \(W_{eff}\):

\[
I_{on(DG)} = W_{eff} Q_{is} v_s \left(1 + \frac{Q_{ib} v_b}{2Q_{is} v_s}\right).
\]

(3.2)

We note that \(v_b\) and \(v_s\) are comparable because of the tendency for velocity saturation along most of the short channel. However, we could get a \(\mu_b > \mu_s\) benefit with velocity overshoot, and hence more contribution of bulk inversion to \(I_{on(DG)}\) via (3.2).

With the substantive bulk inversion charge in the DG FinFET defining \(I_{on(DG)}\) as in (3.2), activation of the top gate, rendering the TG
counterpart, results in only a relatively small increase in the total inversion charge, at the top surface as shown in Fig. 3.5, and in $I_{on}$ as reflected by Figs. 3.2 and 3.3. We confirm then that, indeed, the discrepancies in Fig. 3.3 between the actual and expected $I_{on}$ of the TG FinFET relative to the DG FinFET are mainly a reflection of the significance of the bulk-inversion component of current in (3.1) and (3.2). In fact, this bulk current is the predominant component of $I_{on(DG)}$ in all the DG FinFETs we simulated. Its predominance varies some because the noted field-fringing (relative) benefit to $I_{on(DG)}$ increases with diminishing $R_f$.

Additional simulations reveal that the significant bulk inversion is linked to the undoped, thin body. Because of no significant depletion charge, the electric potential and carrier density in the subthreshold region are uniform throughout the thin body [Tri03a], as exemplified by the 2-D Medici [Med01] simulation results for arbitrary undoped DG and TG [Fos03b] FinFETs in Fig. 3.6. (Virtually the same uniformity obtains for the SG fully depleted (FD) SOI MOSFET with undoped body and thick BOX [Tri03a].) This means that the off-state current in these devices is proportional to the cross-sectional area of the body/channel: $I_{off} \propto h_{Si} w_{Si}$. As the gate voltage ($V_{GS}$) is increased then, this uniformity tends to be maintained, resulting in significant bulk inversion for strong-inversion conditions. The level of the bulk inversion, e.g., $n(x, z = w_{Si}/2)$ which implies
Figure 3.6 Medici-predicted electron density, versus gate voltage, across the fin-body (vs. normalized $z/w_{Si}$) of long-channel DG ($w_{Si} = 20\text{nm}$) and TG (at $x = 0$ for $w_{Si} = h_{Si} = 30\text{nm}$) $n$FinFETs, both with midgap gate; $V_{DS} = 0V$. 
\( Q_{ib}/Q_{is} \) in (3.2), is governed by the electron screening of the surface electric field, via Poisson's equation without ionized dopant charge:

\[
\frac{dE}{dz} = \frac{-qn}{\varepsilon_{Si}},
\]

(3.3)

which is characterized by the Debye length \( L_D \propto \frac{1}{\sqrt{n}} \). It tends to diminish with increasing \( w_{Si} \), but ultimately saturates (at \( \sim 2 \times 10^{18} \text{cm}^{-3} \)), as shown in Fig. 3.7, because \( L_D \) increases as \( n \) decreases. However, it should be noted that for very thick \( w_{Si} \), SCEs might contribute to forming the high \( n(x,z=w_{Si}/2) \).

3.3 Implied Insight of Bulk Inversion

3.3.1 Proper Effective Gate Width

Because of the noted bulk inversion in the undoped DG FinFET, in the off- as well as the on-states, the effective width of the two side fin surfaces, \( 2h_{Si} \), does not properly reflect all the inversion charge and current. The effective gate width should be defined simply as

\[ W_{\text{eff}}(DG) = h_{Si}, \]

(3.4)

with (3.2) modified accordingly. (Note that the effective gate width of the planar SG FD/SOI MOSFET, with bulk inversion, is still the actual gate width \( W_g \), which conveys the \( Q_{ib} \) as well as the \( Q_{is} \) contributions to current.) The gate capacitance is also properly defined by (3.4), i.e., by the area \( L_{\text{eff}}h_{Si} \), as evident in the DG charge characterizations in [Kim01] for asymmetrical- as well as symmetrical-gate devices. However, a proper
Figure 3.7 Davinci-predicted on-state electron density at the middle of the top fin-body surface, at the center of the channel ($y = L_{eff}/2$), of the DG FinFET without the top gate stack of Fig. 3.2, versus the fin width.
effective gate width for the TG FinFET cannot be so directly defined, as evidenced in Fig. 3.3.

3.3.2 Layout Area

We have previously shown [Yan05], based on the commonly defined \( W_{\text{eff}} = 2h_{Si} + w_{Si} \), the gate layout-area inefficiency of TG CMOS relative to DG and FD/SOI CMOS when the undoped TG fin-body dimensions are made comparable to the gate length \( (L_g) \) to ease the fabrication [Doy03]. The bulk inversion noted here clearly worsens this TG inefficiency. For more viable TG CMOS, with taller and thinner fins (and wider \( W_{\text{eff}} \)) as in Sec. 3.2, we further examine the layout-area efficiency, now accounting for bulk inversion, as follows. For a given \( L_g \) and current drive, corresponding to the gate area \( A_{SG} = L_gW_g \) for a planar SG MOSFET (e.g., an FD/SOI MOSFET), the area requirement for the (multi-fin) DG FinFET is \( A_{DG} = L_g[W_gP/(h_{Si}f_{DG})] \), where \( P \) is the pitch and, with reference to (3.4), \( f_{DG} \) is the current-enhancement factor afforded by DG relative to SG when \( h_{Si} = W_g \). Typically, \( f_{DG} > 2 \) [Fos02], but we will assume \( f_{DG} = 2 \) here, which is tantamount to letting \( W_{eff} = 2h_{Si} \) for the DG device. Then for the TG FinFET, we can express \( A_{TG} = L_g[W_gP/W_{eff(TG)}] \), where, phenomenologically, we define

\[
W_{eff(TG)} = 2h_{Si} + w_{Si(\text{eff})};
\]  

(3.5)

because of the bulk inversion, \( w_{Si(\text{eff})} < w_{Si} \) as defined, for \( f_{DG} = 2 \), by the simulation results in Fig. 3.3:
For $R_f = 3$, (3.6) and Fig. 3.3 yield $w_{Si(eff)} = 4.2$nm, much less than the actual $w_{Si} = 13$nm.

The required gate layout areas for the TG and DG FinFETs with $R_f = 3$, relative to $A_{SG}$, are plotted in Fig. 3.8 versus $L_g$; $L_g$ and $P$ were obtained from the 2003 SIA ITRS [Sem03] projections for the HP (high-performance) and LSTP (low-standby power) CMOS applications, and $w_{Si}$ was set to $L_g/2$ ($=L_{eff}/2$) for SCE control. For comparison, we include $A_{TG}/A_{SG}$ that results when $w_{Si(eff)} = w_{Si}$ is assumed, i.e., when bulk inversion is ignored as in [Yan05]. With this assumption the needed TG area is underestimated by about 10% generally for both applications. The actual layout-area ratios, with bulk inversion, show only a minimal benefit of the third gate relative to DG-FinFET CMOS. For the DG technology relative to the planar SG CMOS, the results in Fig. 3.8 are overly pessimistic, showing, for example, >60% more area needed for DG FinFETs in the HP application. Indeed, with $f_{DG} > 2$, which is likely [Fos02], and $R_f > 3$, which is doable, DG-FinFET CMOS can yield significantly better layout-area efficiency than the SG technology [Yan05]. For example, $R_f \to 5$ alone renders DG more area-efficient than SG.

The significance of bulk inversion implies much about nanoscale FinFET characteristics and design. First, the commonly defined $W_{eff}$ is not a valid indicator of relative values of current (and capacitance) in DG and

$$w_{Si(eff)} = 2h_{Si} \left[ \frac{\Delta I_{on}}{I_{on(DG)}} \right].$$

(3.6)
Figure 3.8 Calculated gate layout-area ratios of TG and DG FinFETs with $R_f = h_{Si}/w_{Si} = 3$, relative to the planar SG MOSFET, versus gate length; $L_g (=L_{eff})$ and pitch for the calculations were obtained from the 2003 ITRS [Sem03] projections for HP and LSTP CMOS technologies. The pessimistic area requirements for the DG FinFET resulted from the assumptions of low $R_f$ and $f_{DG} (=2)$.
TG FinFETs; indeed, the additional $I_{on}$ and in fact $I_{off}$, produced by the top gate of the TG device are substantively smaller than what is implied by $W_{eff}$. Second, the top gate is really not needed for moderate $R_f$, which is necessary for good layout efficiency. Third, the gate layout-area advantage of the TG FinFET over the DG counterpart implied by $W_{eff}$ is actually much smaller due to bulk inversion, and the DG area advantage over the simple TG device with $h_{Si} \sim w_{Si} \sim L_g$, noted previously from the $W_{eff}$ perspective [Yan05], is enhanced. Fourth, although we did not consider quantization effects in this study, we surmise that they will actually enhance the bulk-inversion effects because of the deeper inversion-charge centroid in the quantum-mechanical solution [Ge02].

3.4 Summary

Three-dimensional numerical simulations of DG and TG FinFETs having undoped thin bodies have revealed the significance of bulk-inversion current in $I_{on}$, as well as $I_{off}$, and the consequent insignificance of the commonly defined effective gate width in comparisons of DG and TG currents. In fact, we have inferred that the proper $W_{eff}$ for DG FinFETs is $h_{Si}$, which correlates with the total (surface plus bulk) inversion charge; whereas a meaningful $W_{eff}$ cannot be directly defined for TG FinFETs. The new insights revealed herein explain why the DG FinFET provides nearly the same $I_{on}$ as the TG counterpart for fin aspect ratios as small as two, but especially for higher $R_f$ which is desirable and doable. Due to the relatively small increase in $I_{on}$ of TG
FinFETs, over the DG counterparts with moderate $R_f$, the advantage of TG devices in gate layout-area efficiency is not significant. The insights thus further solidify our notion, based initially on $W_{\text{eff}}$-implied TG layout-area inefficiency [Yan05] (and on the fact that a TG FinFET, with a thin top dielectric and moderate $R_f$, is more difficult to fabricate than a DG FinFET [Mat05]), that the third top gate is neither desirable nor beneficial.
CHAPTER 4
MODELING AND SIGNIFICANCE OF FRINGE CAPACITANCE IN NONCLASSICAL CMOS DEVICES WITH GATE-SOURCE/DRAIN UNDERLAP

4.1 Introduction

Nonclassical nanoscale silicon CMOS devices, e.g., double-gate (DG) and single-gate (SG) fully depleted (FD) SOI MOSFETs with undoped ultra-thin bodies (UTBs), should be designed with gate-source/drain (G-S/D) underlap [Tri05a]. The benefits of the underlap include better control of short-channel effects (SCEs) via a gate bias-dependent effective channel length (L_{eff}) [Fos03c, Tri05a], as well as elimination of gate-induced drain leakage (GIDL) [Tan05] and gate-drain/source tunneling currents. The underlap, however, must be optimally designed because it tends to increase the S/D series resistance (R_{SD}) and decrease I_{on} [She03, Tan05]. We show in this chapter, by device and circuit modeling and simulation, that the implied underlap design tradeoff for ultimate CMOS speed is affected significantly by parasitic G-S/D capacitance, i.e., fringe capacitance, in nanoscale devices.

Fringe capacitance in classical MOSFETs, with G-S/D overlap, was modeled some time ago [Shr82], and some modeling was recently reported [Ban05] for DG MOSFETs with underlap. However, the modeling of Bansal et al. [Ban05] focused on a bias-independent outer-fringe
capacitance, and did not address the $V_{GS}$ dependence of it as well as the inner-fringe component, which is quite important in nanoscale devices. Herein, using physical insights derived from numerical device simulations, we develop a complete analytical model for parasitic capacitance in nonclassical devices with G-S/D underlap, which includes both the outer- and inner-fringe components with $V_{GS}$ dependences, as well as a BOX-fringe component in the FD/SOI MOSFET. The new modeling is verified by 2-D numerical device simulations. Further, the model is implemented in our process/physics-based compact model UFDG (Ver. 3.5) [Fos06a], and used in Spice3 simulations to check the benefit of G-S/D underlap in reducing the fringe capacitance and DG CMOS propagation delay. Based on physical insights attained, optimization of the underlap design to effect the best tradeoff between the capacitance and $R_{SD}$ for CMOS speed is exemplified at the $L_g = 18\text{nm}$ technology node of the SIA ITRS [Sem03].

4.2 Physical Insights from Numerical Simulations

4.2.1 Inner and Outer Fringe Capacitance

The dependences of the parasitic capacitance on $V_{GS}$ and on the G-S/D underlap are reflected in Fig. 4.1, which shows low-frequency, low-$V_{DS}$ total gate capacitance ($C_G$) versus $V_{GS}$ predicted by the 2-D device simulator MEDICI [Med04] for an $L_g = 18\text{nm}$ undoped-UTB DG nMOSFET with and without underlap, and with and without a finite gate height ($t_g$). Because of the floating UTB and the negligibly small junction capacitance,
Figure 4.1 MEDICI-predicted low-frequency gate capacitance versus gate voltage for an $L_g = 18\text{nm}$ DG nMOSFET with (graded $N_{SD}(y)$ in 20nm S/D extension with 11nm straggle [Tri05a]) and without (abrupt $N_{SD}(y)$) G-S/D underlap, and with and without finite gate thickness; undoped UTBs with $t_{Si}=14\text{nm}$, $t_{ox}=0.7\text{nm}$, midgap gate.
which is in series with the intrinsic gate-to-body capacitance, the subthreshold $C_G$ is defined exclusively by the extrinsic parasitic G-S/D capacitance, which includes both inner- ($C_{if}$) and outer-fringe ($C_{of}$) components [Ban05]. The same is essentially true for SG FD/SOI MOSFETs with thick BOX [Kim05a]. For the $t_g = 0$ simulations of Fig. 4.1, $C_{of}$ was forced to zero by removal of the spacer dielectric as well as the gate stack from the device domain. Note first then in Fig. 4.1 how significant $C_{if}$ and $C_{of}$ are, relative to the on-state $C_G$. (For the simulations we assumed silicon-dioxide spacers. When they are silicon-nitride, which has about a 1.8x-higher permittivity, $C_{of}$ is even more significant than implied in Fig. 4.1.) For the $t_g = 0$ cases, in which $C_{of} = 0$, note how the G-S/D underlap significantly reduces the subthreshold $C_G$, i.e., $C_{if}$, but makes no difference in strong inversion (at low $V_{DS}$). This reflects the screening of the inner G-S/D fringing electric field by inversion charge, which forces $C_{if}$ (and the BOX-fringe capacitance, as we discuss later) to zero with increasing $V_{GS}$. (At high $V_{DS}$, the reduced inversion charge near the drain could undermine this screening effect, but for nanoscale devices, the carrier velocity saturation tends to keep the inversion charge high enough to make this undermining negligible.) Then, with finite $t_g$, $C_G$ is increased, with and without underlap, for all $V_{GS}$ by $C_{of}$. However, note that $C_{of}$ is smaller in the device with underlap in the subthreshold region, but not in strong inversion. This reflects the shrinking $L_{eff}$ with increasing $V_{GS}$ due
to the decreasing Debye length defined by the strong-inversion charge: \( L_{\text{eff}} \rightarrow L_g \) as \( V_{GS} \) increases [Fos03c].

### 4.2.2 BOX Fringe Capacitance in FD/SOI MOSFETs

More insight is gained from the MEDICI-predicted subthreshold \((V_{GS} = 0) C_G\) versus undoped-UTB thickness \((t_{Si})\) in Fig. 4.2 for FD/SOI nMOSFETs. For the assumed \( t_g = 0 \) \((C_{of} = 0)\), the increasing \( C_G \) with \( t_{Si} \) reflects the dependence of \( C_{if} \) on \( t_{Si} \). However, note for all cases (different \( L_g \), with and without underlap) that \( C_G \) for \( t_{Si} \rightarrow 0 \) remains finite, even though \( C_{if} \) must approach zero. Analogous to the short-channel effect of S/D field fringing in the BOX [Tri03a], which can induce a leakage path near the back surface, we infer that the finite \( C_G \) at \( t_{Si} = 0 \) in Fig. 4.2 is associated with the BOX-fringe capacitance \((C_{bf})\). Note that for increasing \( V_{GS} \), \( C_{bf} \), like \( C_{if} \), will approach zero due to inversion-charge screening.

Based on the physical insights gained from Figs. 4.1 and 4.2, we conclude that there are three basic components of parasitic fringe capacitance in nonclassical devices with G-S/D underlap: \( C_{of}, C_{if}, \) and \( C_{bf} \) as represented in Fig. 4.3, all of which depend on \( V_{GS} \). Actually, \( C_{bf} \) as shown in the figure for SG FD/SOI MOSFETs is an approximation for the G-S/D capacitance supported by the fringing field in the BOX. More exactly, this capacitance (per unit width \( W \)) is a series combination of \( C_{bf} \) (as characterized in the next section), \( C_{ox} = (\varepsilon_{ox}/t_{ox})L_g/2 \), and \( C_b = (\varepsilon_{Si}/t_{Si})L_g/2 \) [Kim05a]. However, for typical thin \( t_{ox} \) and \( t_{Si} \), \( C_{bf} \ll C_{ox}, C_b \), and therefore this combination can be simply characterized as \( C_{bf} \). We note
Figure 4.2 MEDICI-predicted low-frequency subthreshold gate capacitance versus undoped UTB thickness for SG FD/SOI nMOSFETs with two different short gate lengths and assumed flat gate (\( t_g = 0 \), which means no outer-fringe capacitance), with and without G-S/D underlap; \( t_{BOX} = 200\text{nm} \), midgap gate.
Figure 4.3 A schematic diagram of the gate-source/drain structure of a nonclassical (DG or SG FD/SOI) MOSFET, indicating the G-S/D underlap (with effective length $L_{eSD}$) and the three components of the parasitic fringe capacitance; $C_{bf}$ is unique to the FD/SOI device with thick BOX.
that \( C_{of} \) is defined predominantly by the gate sidewalls; the fringe capacitance from the top of the gate is negligible for typical (high) \( t_g \). In strong inversion, \( C_{if} \) and \( C_{bf} \) are effectively screened out, and \( C_{of} \) (with \( L_{eff} \equiv L_g \)) is the main parasitic. As \( L_g \) is scaled, the parasitic capacitance becomes more significant, and hence modeling it is crucial.

### 4.3 Analytical Modeling

To model the parasitic fringe capacitance for SOI-based nanoscale MOSFETs with G-S/D underlap, including DG and SG FD/SOI devices, we extend the modeling in [Shr82] based on the insights gained in Sec. 4.2. The basic fringe-capacitance model stems from that defined by two separated, conducting plates at an angle \( \theta \) as shown in Fig. 4.4. A solution of Laplace’s equation in cylindrical coordinates for the electric potential when a voltage \( V \) is applied as shown yields, via Gauss’s law, the charge \( (Q = CV) \) on the plates and the capacitance (per unit width) it defines [Zah79]:

\[
C = \frac{\varepsilon}{\theta} \ln \left( \frac{r_2}{r_1} \right) \tag{4.1}
\]

where \( \varepsilon \) is the permittivity of the insulator between two plates, and \( r_1, r_2, \) and \( \theta \) are the geometrical parameters defined in Fig. 4.4. The fringing field from the ends of the plates is ignored, assuming the plates are in close proximity [Zah79]. To use (4.1) for the fringe-capacitance components in Fig. 4.3, the basic G-S/D structure is transformed to those of Fig. 4.5, depending on the underlap, or inversion condition: Fig. 4.5(a)
Figure 4.4  Basic two-plate model for fringe capacitance (per unit width in z), with the cylindrical coordinates (r and φ) used in the analysis shown.
Figure 4.5  Schematics of (a) the G-S/D underlap structure for the weak-inversion analysis, with the reduced angle $\beta$ defined after replacing the higher-permittivity silicon with oxide [ Shr82], and (b) the abrupt G-S/D structure with no underlap for the strong-inversion analysis.
approximates the underlap structure in weak inversion, with effective \( L_{eSD} \), where \( L_{eff} \equiv L_g + 2L_{eSD} \) [Fos03c, Tri05a]); and Fig. 4.5(b) shows the effective abrupt G-S/D structure without underlap in strong inversion, with \( L_{eSD} = 0 \), where \( L_{eff} \equiv L_g \) [Fos03c, Tri05a]. (We are neglecting the accumulation condition.) As illustrated in Fig. 4.6 and described in [Tri05a], \( L_{eSD} \) is an effective underlap with an abrupt source/drain-body junction as defined by the actual graded lateral doping profile \( N_{SD}(y) \) with finite straggle \( (\sigma_L) \) in the S/D extension; \( L_{eSD} \) depends on \( \sigma_L \), the extension length \( (L_{ext}) \), and \( t_{Si} \). In Fig. 4.5(a), as was done in [Shr82], the plate-plate angle has been reduced from \( \pi/2 \) to \( \beta \) to effectively account for the silicon permittivity \( (\varepsilon_{Si}) \) being about three-times that of the oxide \( (\varepsilon_{ox}) \), which will be assumed in the use of (4.1):

\[
\beta = \frac{\pi \varepsilon_{ox}}{2\varepsilon_{Si}} \equiv \frac{\pi}{6}. \tag{4.2}
\]

Further, we assume that the spacer dielectric is silicon-dioxide. For silicon-nitride spacers, \( \varepsilon_{ox} \) in our results for \( C_{of} \) (in (4.3) and (4.6)) should be replaced by the permittivity of the nitride.

For UFDG, the \( V_{GS} \) dependence of the fringe capacitance \( (C_f) \) is accounted for by modeling the various components in weak and strong inversion, defining the corresponding charge components (i.e., \( C_fV_{GS/D} \)) and assigning them to the proper terminals of the device, and then linking the models across the moderate-inversion region defined by contour boundaries in \( V_{Gfs}-V_{GbS} \) space [Fos06b]. The links are (“2-D”) cubic-
Figure 4.6 Illustration of how varying the actual lateral doping density profile $N_{SD}(y)$ in the S/D extension changes the weak-inversion effective channel length, as characterized by the effective G-S/D underlap $L_{eSD}$ indicated, which is defined by $\sigma_L$, $L_{ext}$, and $t_{Si}$ as described in [Tri05a].
polynomial spline functions of the two gate voltages for the terminal charges. The UFDG accounting for the $V_{GS}$-dependent $L_{eff}$ [Fos03c, Fos06b, Tri05a] is similarly facilitated by such regional modeling.

### 4.3.1 Weak Inversion

In the weak-inversion region, the outer-fringe capacitance ($C_{ofw}$) can be modeled with (4.1) by defining, from Fig. 4.5(a), $r_1 = \overline{o-f}=\overline{c}=L_{eSD}$, $r_2 = \overline{o-b+bg}=t_{ox}+t_g$, and $\theta = \alpha = \pi/2$. For the inner-fringe capacitance ($C_{if}$) with (4.1), $r_1 = \overline{o-c+cd}=\overline{a+ab}=t_{ox}/\tan\beta+L_{eSD}$, $r_2 = \overline{c+ce}=t_{ox}/\sin\beta+t_{Si}$, and $\theta = \beta$ in (4.2). Thus,

$$C_{ofw} = \frac{2\varepsilon_{ox}}{\pi} \ln \left( \frac{t_r + t_{ox}}{L_{eSD}} \right)$$  \hspace{1cm} (4.3)

and

$$C_{if} = \frac{6\varepsilon_{ox}}{\pi} \ln \left( \frac{t_{ox} + t_{Si}\sin\beta}{t_{ox}\cos\beta + f_{if}L_{eSD}\sin\beta} \right).$$  \hspace{1cm} (4.4)

Note in (4.4) that, because of uncertainty in $L_{eSD}$ (used here for $C_{if}$, as opposed to that used to define $L_{eff}$ [Tri05a]) due to the graded $N_{SD}(y)$ in real devices, we employ the tuning parameter $f_{if}$ by replacing $L_{eSD}$ by $f_{if}L_{eSD}$, which can be thought of as an effective underlap length for $C_{if}$.

From numerical simulations, we find that $f_{if}$ is positive, and generally comparable to but less than unity. To keep the model simple, yet realistic, we have assumed $L_{eSD} \geq t_{ox}$ to get (4.3), and then to get (4.4) that $L_{eSD}$ is less than or equal to $t_{ox}(1-cos\beta)/\sin\beta + t_{Si} \equiv 0.27t_{ox}+t_{Si}$, since otherwise $\overline{de}$ in Fig. 4.5(a) is nonexistent (i.e., $\overline{o-a+ab} > \overline{o-c+ce}$) and $C_{if}$ is undefinable.
via (4.1). These assumptions do not limit the model utility since typically $L_{eSD} > t_{ox}$ and $L_{eSD} < t_{Si}$ [Tri05a]. In Fig. 4.5(a), we also assumed that the electric field from the region $\overline{bf}$, including fringing field, does nothing to $C_{ofw}$ and $C_{if}$ since the gate generally suppresses accumulation charge in the underlap region. However, when $L_{eSD}$ is large ($> \sim 4$nm, which is generally too long due to high $R_{SD}$ [Fos03c]), the effect from the region $\overline{bf}$ is not totally negligible, and thus the model tends to underestimate $C_{ofw}$ a bit.

For DG devices, the total parasitic G-S/D capacitance in the weak-inversion region is approximated by parallel combination of (4.3) and (4.4), applied for both gates coupled to both the source and the drain. For SG FD/SOI devices, there is only the front gate, but the BOX-fringe capacitance must be accounted for. Following our discussion in Sec. 4.2, still using (4.1), we can model $C_{bf}$ in Fig. 4.3 as indicated in Fig. 4.7, where the two plates can be defined by $\overline{ab}$ and $\overline{cd}$. Thus, by defining $r_1 = \overline{bo} = \overline{oc} = L_{eSD}/2$, $r_2 = \overline{ab} + \overline{bo} = \overline{oc} + \overline{cd} = L_{eSD}/2 + L_g/2$, and $\theta = \gamma = \pi$, we get

$$C_{bf} = \frac{\varepsilon_{ox}}{\pi} \ln \left( 1 + \frac{L_g}{L_{eSD}} \right). \quad (4.5)$$

For weak inversion, as we noted previously, we can assume that (4.5) couples the source/drain and gate directly, ignoring any $t_{Si}$ dependence, since a series combination of $C_{bf}$, $C_{ox}$, and $C_b$ can be simply characterized as $C_{bf}$ due to $C_{bf} << C_{ox}$, $C_b$ for typical $t_{ox}$ and $t_{Si}$. Hence, for SG FD/SOI MOSFETs, the total parasitic G-S/D capacitance in the weak-inversion
Figure 4.7 Schematic of the SG FD/SOI MOSFET with G-S/D underlap, showing how the BOX-fringe capacitance is modeled.
region is approximated by the parallel combination of (4.3), (4.4), and (4.5), applied to both the source and the drain.

4.3.2. Strong Inversion

As explained in Sec. 4.2, the outer-fringe capacitance in the strong-inversion region \(C_{ofs}\) should always be defined by an effective abrupt gate-source/drain structure with \(L_{eSD} = 0\) as shown in Fig. 4.5(b). This means that \(C_{ofs}\) can be expressed, as in [Shr82], with (4.1) by defining \(r_1 = o_a = o_b = t_{ox}\) and \(r_2 = o_b + b_c = t_{ox} + t_g\) with \(\theta = \alpha = \pi/2\). Thus,

\[
C_{ofs} = \frac{2e_{ox}}{\pi} \ln \left(1 + \frac{t_g}{t_{ox}}\right). \tag{4.6}
\]

Note here that the fringing-field effects due to region \(o_a\) in Fig. 4.5(b), which was accounted for in [Shr82] quasi-empirically, is not significant because \(t_{ox}\) is ultra-thin in nanoscale MOSFETs and the inversion charge tends to obviate any accumulation charge in the \(o_a\) region.

So, for strong inversion, where \(C_{if}\) and \(C_{bf}\) are negated by inversion-charge screening, the parasitic capacitance, for both DG and SG FD/SOI MOSFETs, is given by (4.6), applied to the gate(s) coupled to both the source and the drain.

4.4. Model Verification

For model verification, an \(L_g = 25\)nm undoped-UTB SG FD/SOI nMOSFET with midgap gate, and \(t_{ox} = 1\)nm, \(t_{Si} = 6\)nm, \(t_g = 20\)nm, and \(t_{BOX} = 200\)nm, along with the G-S/D underlap defined by \(N_{SD}(y)\) in a 30nm S/D extension with a 15nm straggle (which yields \(L_{eSD} = 3.4\)nm for \(L_{eff}\) in
UFDG [Tri05a]) is considered first. For this device in the weak-inversion region (at $V_{GS} = 0V$), MEDICI predicts for the total S/D (i.e., S or D) fringe capacitance, $C_f = C_{ofw} + C_{if} + C_{bf} = 0.111fF/\mu m$, where $C_{ofw} = 0.039fF/\mu m$, $C_{if} = 0.047fF/\mu m$, and $C_{bf} = 0.025fF/\mu m$. In the strong-inversion region, $C_f = C_{ofs} = 0.065fF/\mu m$ is predicted. For the same device, our analytical model predicts, in weak inversion, $C_f = 0.110fF/\mu m$, which is the sum of $C_{ofw} = 0.04fF/\mu m$, $C_{if} = 0.047fF/\mu m$ with $f_{if} = 0.64$ in (4.4), and $C_{bf} = 0.023fF/\mu m$, while $C_{ofs} = 0.067fF/\mu m$ in strong inversion. The model predictions are very good.

For an $L_g = 18$nm undoped-UTB DG nMOSFET with midgap gate, and $t_{ox} = 0.7$nm, $t_{Si} = 14$nm, and $t_g = 18$nm, along with the G-S/D underlap defined by $N_{SD}(y)$ in a 20nm S/D extension with an 11nm straggle ($L_{eSD} = 4.0$nm for $L_{eff}$ in UFDG [Tri05a]), we get $C_f = 0.125fF/\mu m$ in the weak-inversion region (at $V_{GS} = 0V$) from MEDICI, comprising $C_{ofw} = 0.046fF/\mu m$ and $C_{if} = 0.079fF/\mu m$, while our model predicts $C_f = 0.113fF/\mu m$, with $C_{ofw} = 0.034fF/\mu m$ and $C_{if} = 0.079fF/\mu m$ with $f_{if} = 0.85$. In the strong-inversion region, MEDICI predicts $C_{ofs} = 0.072fF/\mu m$, which is also predicted by our model. Again, the model predictions are very good, except for $C_{ofw}$, for which there is a 26% error. This relatively large error, especially for DG devices, comes from the ignored $bf$ region in Fig. 4.5(a), which might contribute to $C_{ofw}$ when $L_{eS}$ is relatively large. Nonetheless, our model overall agrees very well with the 2-D simulation results from MEDICI, including additional ones for other SG FD/SOI (with $L_g = 13$nm
and $L_{eSD} = 2.7\text{nm}$) and DG ($L_g = 7\text{nm}$ and $L_{eSD} = 2.0\text{nm}$) MOSFETs, showing generally $<15\%$ errors. We note that the tuning parameter $f_{if}$ in (4.4) tends to increase and approach unity with decreasing $L_{eSD}$, especially for DG devices. For short underlaps, the source/drain-body doping profile tends to be more abrupt, removing uncertainty in $L_{eSD}$.

4.5 Model Implementation in UFDG (Ver. 3.5)

Now, we implement the analytical model for parasitic fringe capacitance in UFDG (Ver. 3.5) [Fos06a, Fos06b], with the $V_{GS}$ dependences accounted for. The process/physics basis of UFDG, with rigorous accountings for SCEs (via a 2-D solution of Poisson’s equation in the UTB), quantization (QM) effects (via a self-consistent solution of the Poisson and Schrödinger equations in the UTB [Ge02] that describes the bulk inversion [Kim05b]), and carrier transport in the UTB/channel (via a QM-based mobility model [Fos06b] with carrier temperature-dependent velocity overshoot [Ge01] and carrier injection-velocity defined ballistic-limit current [Fos06b]), makes it quasi-predictive and hence useful for projecting nonclassical nanoscale device/circuit performance. The implementation was facilitated by the regional analyses for weak and strong inversion used in UFDG, which are linked by (“2-D”) $V_{Gfs}$- and $V_{GbS}$-based cubic splines for charge (and current) across the moderate-inversion region defined by contour boundaries in $V_{Gfs}$-$V_{GbS}$ space [Fos06b]. The UFDG accounting for the $V_{GS}$-dependent $L_{eff}$ [Fos03c, Tri05a] was similarly facilitated by the noted regional modeling.
Model implementation has been done as follows. At first, we use a flag (CFF) to turn on (CFF = 1) and off (CFF = 0) the parasitic fringe capacitance model. Then, \( t_{ox} + t_g \) is compared to \( L_{eSD} \), and if it is smaller than \( L_{eSD} \) or \( t_g \) is zero, \( C_{of} \) is set to be zero since \( C_{of} \) should be zero, while (4.3) predicts a negative solution. When it is larger than \( L_{eSD} \), \( C_{of} \) is calculated with (4.3). After the \( C_{of} \) calculation, \( C_{if} \) calculation will follow. As mentioned in Sec. 4.3, the total parasitic G-S/D capacitance for DG devices in the weak-inversion region is approximated by these two components, i.e., \( C_{of} \) and \( C_{if} \), applied for both gates coupled to both the source and the drain. Here, it should be noted that the model is accurate for a certain range as defined in Sec. 4.3, and thus \( L_{eSD} \) should be set to a proper constant value when it is out of the defined range. So, for \( L_{eSD} < t_{ox} \), \( L_{eSD} \) is set to \( t_{ox} \) for \( C_{of} \) calculations, while, for \( L_{eSD} < 0.27t_{ox} \), \( L_{eSD} \) is defined to \( 0.27t_{ox} \) for \( C_{if} \) calculations. Therefore, for abrupt case without underlap, the model calculates \( C_{of} \) and \( C_{if} \) with \( L_{eSD} = t_{ox} \) and \( L_{eSD} = 0.27t_{ox} \), respectively. These constant values are consistent with those in [Shr82]. Also, when \( L_{eSD} > 0.27t_{ox} + t_{Si} \), \( C_{if} \) is set to zero in UFDG, because the model is not defined in this region. For SG FD/SOI devices, the BOX-fringe capacitance is included in the total parasitic capacitance as explained. Since the BOX-fringe capacitance is modeled with the assumption that two conducting plates are placed separately, \( L_{eSD} \) cannot be zero. Therefore, in UFDG, the minimum value of \( L_{eSD} \) is set to 1.2nm, which is empirically obtained from the numerical simulations.
4.6 Model Applications

We now use UFDG/Spice3 to access the effect of $C_f$ on DG CMOS speed and to check the benefit of an optimal G-S/D underlap on DG CMOS speed, as well as show more verification of our fringe capacitance model. To do this, we consider the HP45 technology node [Sem03] with $L_g = 18\text{nm}$ (and silicon-dioxide spacers). We assume DG MOSFETs (e.g., FinFETs [Hua99]) with a midgap gate with $t_g = 18\text{nm}$, and undoped UTBs with $t_{Si} = L_g/2 = 9\text{nm}$. (Actually, $t_{Si} = L_{eff}/2$ gives good SCE control [Yan05], so, for devices with underlap, we are using thinner $t_{Si}$ than is needed. We do this because we want to compare the underlap-device performance with that of a well-tempered device without underlap, i.e., one with an abrupt S/D-body junction for which $L_{eff} = L_g$.) Throughout the study, we generally use the 2003 ITRS [Sem03] as a reference, except for $t_{ox}$ and $R_{SD}$ specifications. The gate leakage current can be controlled with thicker $t_{ox}$, e.g., 1.0nm instead of 0.7nm given in the ITRS, enabling a pragmatic yet optimal DG CMOS [Fos04b]. The G-S/D underlap region implies, compared to the abrupt junction, higher $R_{SD}$, which must include a component ($\Delta R_{SD}$) defined by $N_{SD}(y)$. Note here that a more rigorous design optimization study of the underlap will be given in chapter 5.

In Fig. 4.8, UFDG-predicted gate capacitances versus $V_{GS}$ are compared with those from MEDICI simulations for the DG nMOSFET, with and without (abrupt $N_{SD}(y)$) G-S/D underlap. Note here that for the abrupt $N_{SD}(y)$ with $L_{eSD} = 0$, (4.3) and (4.4) do not apply directly. However,
Figure 4.8 UFDG- and MEDICI-predicted gate capacitance versus gate voltage for the $L_g = 18\text{nm}$ DG nMOSFET ($t_{Si} = 9\text{nm}$, $t_{ox} = 1\text{nm}$, $t_g = 18\text{nm}$, midgap gate), with and without (abrupt $N_{SD}(y)$) G-S/D underlap. For the near-optimal underlap, $L_{eSD} = 3.4\text{nm}$ is defined by a graded $N_{SD}(y)$ in a 15nm S/D extension with 9nm straggle [Tri05a].
for cases like this, UFDG assumes finite values for $L_{eSD}$ that make (4.3) and (4.4) reasonably consistent with [Shr82]. For the underlap case, a near-optimal $L_{eSD} = 3.4\text{nm}$, with regard to the $C_f$ vs. $R_{SD}$ tradeoff, was obtained from $N_{SD}(y)$ in a 15nm S/D extension with a 9nm straggle [Tri05a]. As can be seen in the figure then, with $f_i$ tuned to give good subthreshold $C_G$ matches, the predicted results are in good agreement with those from MEDICI for both device structures, again showing the benefit of the underlap in reducing $C_G$ in the weak-inversion region. However, as shown by the UFDG (with $R_{SD}$ tuned to match $I_{on}$) and MEDICI current-voltage predictions in Fig. 4.9, $I_{on}$ is slightly lowered by the underlap due to higher $R_{SD}$. (The QM and velocity overshoot options were not used here because these effects are not modeled well in MEDICI.) Note also the substantive reduction in $I_{off}$ afforded by the underlap, which is related to $L_{eff} > L_g$ [Fos03c, Tri05a]. There is indeed an underlap-design tradeoff regarding $I_{off}$ (or $L_{eff}$), $I_{on}$ (or $R_{SD}$), and $C_G$ (or speed).

For CMOS speed projections and optimal-underlap study, we must ensure that $\Delta R_{SD}$ is correlated properly with $L_{eff}$ and $C_f$, all of which depend on $N_{SD}(y)$ in the S/D extension. We assumed that $R_{SD}$ is pragmatic and constant ($=120\Omega \cdot \mu m + \Delta R_{SD}$) in the strong-inversion region, with $\Delta R_{SD}$ evaluated from the difference between the MEDICI-predicted $I_{on}$ for the underlap (with $L_{eSD}$ defined as noted) and the abrupt-$N_{SD}(y)$ devices, as noted with reference to Fig. 4.9. Then, with the total $R_{SD}$ defined, and $f_{if}$ evaluated as noted with reference to Fig. 4.8, we use UFDG (with the
Figure 4.9 UFDG- and MEDICI-predicted current-voltage characteristics of the $L_g = 18\text{nm}$ DG nMOSFET in Fig. 4.7, with and without the G-S/D underlap.
velocity overshoot [Ge01] and quantization [Ge02] options turned on) in Spice3 to simulate a 9-stage unloaded DG CMOS-inverter ring oscillator. The UFDG/Spice3-predicted propagation delays for different $L_g = 18\text{nm}$ device designs are plotted versus supply voltage ($V_{DD}$) in Fig. 4. 10. For comparison, results for a worst-case abrupt $N_{SD}(y)$ with G-S/D overlap (defined as 10% of $L_g$) capacitance, as in classical MOSFETs, are included in the figure. For this case, $C_f$ is modeled as discussed for the abrupt-$N_{SD}(y)$ device of Fig. 4. 8, and an overlap capacitance equal to $(\varepsilon_{ox}/t_{ox})W(0.1L_g)$ is assumed at the source and drain. Note that such design is ideal when only $I_{on}$ or $R_{SD}$ is considered [Tri05a], but its speed is much slower (41% longer delay) than that for the same device structure (i.e., $t_{ox} = 1\text{nm}$ and $t_{Si} = 9\text{nm}$) with near-optimal underlap. Even without any overlap capacitance, which is not realistic, the abrupt-$N_{SD}(y)$ devices are slower (5% longer delay) than those with underlap. Indeed then, the reduction of $C_G$ afforded by well-tempered underlap translates to faster CMOS speed.

We now explore optimization of the underlap design. As noted in Fig. 4. 9, $I_{off}$ of the device with underlap is much lower than that of the abrupt-$N_{SD}(y)$ device because of the better SCE control. We can thus consider increasing $t_{Si}$ and/or $t_{ox}$ to lower the threshold voltage (via enhanced SCEs and less quantization [Tri03a]) and make $I_{off}$ roughly equal to that of the abrupt-$N_{SD}(y)$ device. Increasing $t_{Si}$ alone may increase $I_{on}$ a bit, but it also increases $C_{if}$ as indicated in (4.4) and as
Figure 4.10 UFDS/Spice3-predicted propagation delays versus supply voltage of 9-stage unloaded DG CMOS-inverter ring oscillators for five different variations of the \(L_g = 18\text{nm}\) device design of Figs. 4.8 and 4.9, with and without the G-S/D underlap. For the worst-case design, a G-S/D overlap of 10% of \(L_g\) was assumed. For the thick \(t_{ox} = 1.5\text{nm}\) device design, the \(V_{DD} = 1.0\text{V}\) delay predicted with the G-S/D fringe capacitance completely removed is plotted as well.
Figure 4.11 UFDG- and MEDICI-predicted gate capacitance versus gate voltage for the $L_g = 18$nm DG nMOSFET with the G-S/D underlap, for varying UTB and oxide thicknesses.
shown by the UFDG and MEDICI simulation results in Fig. 4. 11. Thus, it
does not decrease the delay much as shown in Fig. 4. 10. In fact, the speed
is slightly degraded, except for low $V_{\text{DD}}$ where the $I_{\text{on}}$ increase is more
significant. (Nonetheless, this thicker $t_{\text{Si}} = 12\text{nm} \equiv L_{\text{eff}}/2$ is more
pragmatic technologically, still yielding good SCE control and speed
performance.) Increasing $t_{\text{ox}}$ yields a more pragmatic benefit. For $t_{\text{ox}} =
1.5\text{nm}$, the delay is actually a bit shorter, as seen in Fig. 4. 10, because
both the intrinsic gate capacitance and the parasitic fringe capacitance
(mainly $C_{\text{of}}$ in strong inversion) are reduced with increasing $t_{\text{ox}}$, as
reflected in Fig. 4. 11, while the channel current and $I_{\text{on}}$ decrease at a
lesser rate because of bulk inversion and mobility enhancement [Kim05b].
However, the main benefit of thicker $t_{\text{ox}}$ is not enhanced speed, but
restricted gate tunneling current and avoidance of a high-$k$ dielectric,
without any speed degradation. For such pragmatic design, Fig. 4. 10
shows that the combination of thicker $t_{\text{ox}}$ with near-optimal G-S/D
underlap yields 32% improvement in the CMOS speed at $V_{\text{DD}} = 1.0\text{V}$
compared to that of the abrupt-$N_{\text{SD}}(y)$ design with typical G-S/D overlap;
it is even 9% faster than that of the ideal abrupt-$N_{\text{SD}}(y)$ design without
the overlap. (And, it could be made more pragmatic by using thicker $t_{\text{Si}}$ as
we have intimated.)

The impact of the parasitic fringe capacitance on the CMOS
speed is severe. To emphasize this finding, we include in Fig. 4. 10 the
UFDG/Spice3-predicted ring-oscillator delay at $V_{\text{DD}} = 1.0\text{V}$ for the
pragmatic $t_{\text{ox}} = 1.5\text{nm}$ device design, but with $C_f$ completely removed. The result is dramatic. The delay is reduced from $2.7\text{ps}$ to $0.7\text{ps}$, or by about a factor of four! This result, which would be larger for common silicon-nitride spacers, shows that typical G-S/D fringe capacitance in nanoscale DG CMOS devices, even with optimal G-S/D underlap, plays a predominant role in limiting speed.

4.7 Summary

Using 2-D numerical device simulations, we showed that the parasitic fringe capacitances in nonclassical nanoscale MOSFETs, e.g., DG FinFETs, are significant, with important $V_{GS}$ dependences due to the gate-source/drain underlap that in fact reduces the capacitance. With physical insights from the device simulations, we developed an analytical model for the parasitic capacitance, including inner- and outer-fringe components, and a BOX-fringe component for FD/SOI MOSFETs, all with dependences on $V_{GS}$ and on the underlap structure. The model was verified generally by the numerical simulations, and implemented in our process/physics-based compact model (UFDG-3.5). With UFDG in Spice3, we showed, via ring-oscillator simulations, that reducing the parasitic capacitance via optimal underlap design can be quite effective in improving nanoscale DG CMOS speed, which is basically defined by a tradeoff regarding the capacitance and source/drain series resistance. Further, we showed that, for a given underlap structure, increasing the UTB thickness tends to slightly degrade the device speed due to the
increased inner-fringe capacitance in the weak-inversion region (but still could yield a good pragmatic design). However, increasing the gate-oxide thickness, with near-optimal underlap, can give a pragmatically improved DG CMOS design that avoids gate current and high-k dielectric, without any speed degradation; if fact, we predicted that the speed can actually be enhanced a bit. Such a pragmatic design is possible because increasing $t_{\text{ox}}$ reduces both the parasitic fringe capacitance and the intrinsic gate capacitance, while decreasing the channel current and $I_{\text{on}}$ less because of bulk inversion and mobility enhancement.

Nonetheless, we stress the severity of the G-S/D fringe-capacitance effect on speed shown by our simulations. We found that this parasitic capacitance is predominant in limiting nanoscale DG CMOS speed, even when moderated by an optimal G-S/D underlap. (This statement applies to classical CMOS, without underlap, as well.) Indeed, parasitic capacitance, as well as series resistance, are crucial issues in the design of nanoscale CMOS.
CHAPTER 5
DOUBLE-GATE FINFETS WITH GATE-SOURCE/DRAIN UNDERLAP: APPLICATIONS ON SRAM CELL AND DESIGN OPTIMIZATION FOR DEVICE SPEED

5.1 Introduction

Double-gate (DG) FinFETs with undoped ultra-thin bodies (UTBs) are very attractive for scaled CMOS mainly due to their excellent suppression of short-channel effects (SCEs), high on-state versus off-state current ratio ($I_{on}/I_{off}$), and elimination of threshold voltage ($V_t$) variations caused by statistical dopant fluctuation effects. Higher carrier mobility, which comes from smaller transverse electric field and negligible impurity scattering in the undoped UTBs, and much smaller parasitic junction capacitance are the additional benefits of DG FinFETs. However, with the ultimate limit of the UTB, i.e., $\equiv 5$nm [Tri03a] due to severe quantization effects and technological difficulties, DG FinFET scaling to and beyond the HP25 node with the physical gate length ($L_g$) of 10nm [Sem05] seems to be extremely difficult since the fin width ($w_{Si}$) required for SCE control is $w_{Si} \equiv L_{eff}/2$ [Yan05] if high-k gate dielectric is not viable. Thus, for further gate length scaling to and beyond 10nm, DG FinFETs have to be designed with gate-source/drain (G-S/D) underlap [Tri05a]. Even for the $L_g > 10$nm regime or/and when a reliable high-k gate dielectric is developed, the underlap structure should be quite useful in the device
design for effecting an optimal SCEs versus \( I_{\text{on}} \) trade-off [Kra06, Lim05, Tri05a].

The benefit of an underlap structure in the DG FinFET should be most useful for SRAM applications. This is because the read static noise margin (read-SNM) and write-margin are not defined by the absolute value of \( I_{\text{on}} \), but by \( V_t \) and the relative strength of \( I_{\text{on}} \) among the transistors in SRAM cell. Note here that \( V_t \) can be easily increased by SCE control, with some degradation of \( I_{\text{on}} \), via the effective channel length (\( L_{\text{eff}} \)) modulation in the weak-inversion region [Fos03c]. Also, note that high \( V_t \) tends to give large read-SNM and write-margin based on the large inverter trip point, and small cell leakage currents or standby power due to small \( I_{\text{off}} \) [Guo05]. On the other hand, for the device speed issue, we confirmed that the optimally designed underlap [Kim06] can reduce the propagation delay by limiting the fringe capacitance (\( C_f \)) in weak inversion. However, this design approach is virtually based on the trade-off between \( I_{\text{on}} \) and the parasitic capacitance in weak inversion. Therefore, with regard to the device speed, broader study about the underlap optimization is needed.

In this chapter, we first explore SRAM cell design and scaling via DG FinFETs with G-S/D underlap. For this study, DG FinFETs with the underlap are first characterized in terms of \( V_t \) with various extension length (\( L_{\text{ext}} \)), straggle (\( \sigma_L \)), and \( w_{\text{Si}} \) via 2-D numerical [Med04] and analytical simulations [Fos06a]. The relationship between \( V_t \) and read-SNM is verified to define an optimal SRAM cell, for the HP45 node with
$L_g = 18\text{nm}$ [Sem05], with large read-SNM as well as large write-margin and good immunity to process variations of $L_{\text{ext}}$ and $\sigma_L$. Then, a scalability study of the DG FinFET-based SRAM cell, with and without the G-S/D underlap, is done. Finally, based on the insight gained from $V_t$ shift and $I_{\text{on}}$ variation caused by $\sigma_L$ changes, we optimally design DG FinFETs with relatively high $I_{\text{on}}$ to improve the device speed.

5.2 DG FinFET without Underlap

For HP45 ($L_g = 18\text{nm}$) applications [Sem05], we first characterize the pragmatic DG FinFET with undoped UTBs, which has the equivalent oxide thickness (EOT) of 1nm and $w_{\text{Si}}$ of 9nm. Note here that EOT = 1nm, instead of <0.7nm given in ITRS [Sem05], and $w_{\text{Si}} = L_{\text{eff}}/2 = 9\text{nm}$ [Yan05] are essential to control the gate leakage current [Yan04] and SCEs of a DG FinFET with abrupt S/D-extension doping, $N_{\text{SD}}(y)$, respectively. For this DG FinFET design with a midgap metal gate, UFDG [Fos06a], with the aid of Medici [Med04], predicts $I_{\text{off}}$ of 2.96nA/$\mu\text{m}$ and $I_{\text{on}}$ of 1.26mA/$\mu\text{m}$. Here, the predicted $I_{\text{off}}$ is about two-orders of magnitude lower than the limit in ITRS [Sem05], mainly due to well-controlled SCEs via thin $w_{\text{Si}}$ and the midgap gate. Also, the subthreshold swing (S) is predicted to be less than 90mV/dec due also to the thin $w_{\text{Si}}$. On the other hand, the predicted $I_{\text{on}}$ does not meet the current limit (2.05mA/$\mu\text{m}$) projected by ITRS [Sem05]. This is mainly because of the assumed, thick $t_{\text{ox}}$ to control gate leakage current, and the midgap gate. In addition, $I_{\text{on}}$ of the designed device is further degraded by high S/D series resistance ($R_{\text{SD}} = 85\Omega\cdot\mu\text{m}$), which is currently viable, but much larger than the
specified value ($R_{SD} = 53\Omega \cdot \mu m$) at the HP45 node [Sem05], and the ballistic limit [Tri05b], which might not be accounted for in ITRS $I_{on}$ projections. Unfortunately, without the viable high-k gate dielectric and the technology reducing $R_{SD}$ further, this sacrifice of $I_{on}$ is not avoidable.

Nevertheless, the use of this pragmatic DG FinFET design for SRAM applications is not limited since, as mentioned, the relative strength of currents among transistors in SRAM cell is much more important than their absolute values. In addition, the predicted $V_t$ of the mentioned DG FinFET design is relatively high ($V_t \equiv 0.29V$ when it is defined via $I_{DS}(V_{GS} = V_t) = 10^{-7}W_g/L_g$ (A)) due to the assumed midgap gate and well-controlled SCEs. Therefore, we can infer here that this DG FinFET design is good enough for SRAM applications. Indeed, UFDG/Spice3 predicts relatively good read-SNM (177mV) and write margin (350mV) for 1V supply voltage, which are consistent with the results for the high-$V_t$ DG FinFET designs presented in [Guo05]. Thus, for SRAM study, this pragmatic DG FinFET design without underlap is used as a reference, and compared to other DG FinFET designs with underlap.

For the device speed issue, it seems to be extremely difficult to meet the CMOS speed limit in ITRS due to the mentioned $I_{on}$ degradation, which is inevitable. Nevertheless, it is worthy to study the design optimization of the underlap to improve the device speed of DG FinFETs. So, the CMOS speed issue of DG FinFETs is included in Sec. 5.5 to show the speed superiority of the underlapped devices over the abrupt
counterparts, and provide an optimal design approach in using the underlap.

5.3 Threshold Voltage Modulation by Underlap

Now, to increase $V_t$ of the DG FinFET with the fixed gate electrode, i.e., a midgap gate, we employ the underlap structure. The underlap is defined by $L_{ext}$ and $\sigma_L$ [Tri05a], which affect the device characteristics via the effective underlap ($L_{esD}$, or parameters $L_{es}/L_{eD}$ in UFDG). Since, in the weak-inversion region, $L_{eff}$ is defined by $L_g + 2L_{esD}$ [Fos03c], we can simply control SCEs by defining the underlap, and thus increase $V_t$ of the given device as well as decreasing $I_{off}$, while $I_{on}$ is sacrificed some by the increased S/D series resistance ($\Delta R_{SD}$). To check this effect, the previous DG FinFET design with abrupt $N_{SD}(y)$, i.e., the device with $w_{Si} = 9\text{nm}$ and $t_{ox} = 1\text{nm}$, is modified to have the underlap defined by various $L_{ext}$ and $\sigma_L$. Then, the device characteristics are compared to those of the reference design. Here, note that for SRAM applications the DG FinFET design is first oriented to make the device have high $V_t$ rather than high $I_{on}$, since high $V_t$ is more beneficial for high read-SNM and write margin as well as low standby power. Then, based on the insights gained from $V_t$ study, we optimize DG FinFETs with regard to $I_{on}$ and $C_G$ in weak inversion to improve the device speed.

For the study of DG FinFETs with underlap, and of the SRAM cell they constitute, we use UFDG and Medici in concert. For a particular device with $N_{SD}(y)$ defined by $L_{ext}$ and $\sigma_L$, we first simulate it with Medici, predicting current-voltage characteristics, SCEs, $I_{on}$, etc. (The QM and
velocity overshoot options were not used since these effects are not modeled well in Medici.) Then we calibrate UFDG to the Medici predictions, thereby inferring \( L_{eSD} \), \( L_{eff} \), and \( \Delta R_{SD} \). Here, \( \Delta R_{SD} \) is evaluated from the difference between the Medici-predicted \( I_{on} \) for the underlap and the abrupt-\( N_{SD}(y) \) device. With the FinFET model card thereby defined, UFDG/Spice3 with QM [Ge02] and velocity overshoot model [Ge01] turned on is finally used for the needed device/circuit simulations. Note here that to account for the S/D dopant effects on \( V_t \), the work-function of the gate electrode is accordingly modulated since UFDG does not have a proper model or model parameter to project this effect.

UFDG-predicted \( V_t \) dependence on \( L_{eff} \) (or \( L_{eSD} \)) for different \( L_{ext} \) along with various \( \sigma_L \) is represented in Fig. 5.1. From this simulation result, we learn two things: One is that even though shorter \( L_{ext} \) shows a bit better immunity to \( \Delta L_{eff} \) (\( \Delta L_{eSD} \)) and higher \( V_t \), the various combinations of \( L_{ext} \) and \( \sigma_L \) for the same \( L_{eSD} \) do not affect \( V_t \) much when \( L_{eSD} \) is not too short. The other one is that \( V_t \) can be simply increased by increasing \( L_{eSD} \) via using shorter \( \sigma_L \) for a given \( L_{ext} \). An interesting result here is that \( V_t \) decreases faster for shorter \( L_{eSD} \) and can be even smaller than that of the abrupt \( N_{SD}(y) \) device. This is mainly due to S/D dopants that diffuse into the channel, which happens when long \( \sigma_L \) is used to get short \( L_{eSD} \) for a given \( L_{ext} \). For long symmetrical double-gate (SDG) devices, the threshold voltage is represented by [Tri05b]
Figure 5.1  UFDG-predicted threshold voltage ($V_t$) versus the effective channel length ($L_{\text{eff}}$); $L_{\text{eff}}$ is defined by $L_g (=18\text{nm}) + 2L_{\text{eSD}}$, where different $L_{\text{eSD}}$ can be obtained by the various combination of $L_{\text{ext}}$ and $\sigma_L$; the dotted line shows $V_t$ of the DG FinFET with the abrupt $N_{SD(y)}$ doping profile; $L_{\text{eSD}}$ increases with decreasing $\sigma_L$ for a given $L_{\text{ext}}$. 

<table>
<thead>
<tr>
<th>$L_{\text{ext}}$</th>
<th>$\sigma_L$</th>
<th>$V_t$ (0.295V) for abrupt $N_{SD(y)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>18nm</td>
<td>8-12nm</td>
<td></td>
</tr>
<tr>
<td>15nm</td>
<td>6-10nm</td>
<td></td>
</tr>
<tr>
<td>12nm</td>
<td>4-8nm</td>
<td></td>
</tr>
</tbody>
</table>
where $\Phi_{MS}$ is the work-function difference between gate and body, $Q_B$ is depletion charge density defined by the dopants in the channel ($N_B$) via $qN_BT_{Si}$, and $\phi_c$ is the surface potential at $V_t$ defined by $(kT/q)\ln(10^{11} \text{cm}^{-2}/T_{Si}n_i)$. S/D dopants in the channel can contribute to reducing $V_t$ by the last term in (5.1), i.e., $Q_B/C_{ox}$. Indeed, for $L_{ext} = 18\text{nm}$, when $\sigma_L$ is increased from $8\text{nm}$ ($L_{eSD} = 4.8\text{nm}$) to $12\text{nm}$ ($L_{eSD} = 2.4\text{nm}$), $V_t$ is reduced from $0.40\text{V}$ to $0.28\text{V}$, i.e., total $\Delta V_t = 0.12\text{V}$ as reflected in Fig. 5.2. This decrease is much larger than SCE-governed $\Delta V_t(SCE)$ [Tri03b] ($\approx 0.07\text{V}$), i.e.,

$$
\Delta V_{t} \approx 0.05\text{V},
$$

where $\phi_F$ is the Fermi potential, $\phi_c$ is the surface potential defined in (5.1), and $\lambda_{(SDG)} = (T_{Si}/2)[0.5(1+12T_{ox}/T_{Si})]^{1/2}$. Thus, the additional $V_t$ shift ($\approx 0.05\text{V}$) in Fig. 5.2 should be explained by $Q_B$ in (5.1). Therefore, we conclude that the general $V_t$ reduction is not defined only by SCEs governed $L_{eff}$, but also by $Q_B/C_{ox}$, especially when $\sigma_L$ is long to get short $L_{eSD}$.

So, to effectively increase $V_t$ and thus utilize the underlap for SRAM applications, we have to employ the underlap defined by long $L_{eSD}$ via short $L_{ext}$ and $\sigma_L$. However, long $L_{eSD}$ tends to degrade $I_{on}$ due to the increased $\Delta R_{SD}$ as shown in Fig. 5.3. In addition, the access time in SRAM
Figure 5.2 Threshold voltage reduction ($\Delta V_t$) versus the effective channel length ($L_{\text{eff}}$) with $L_{\text{eff}} = 27.8\text{nm}$ as the reference; the dotted line shows $\Delta V_t$ caused only by SCEs expressed in (5.2).
Figure 5.3 UFDG-predicted on-state current ($I_{on}$) and the increased source/drain series resistance ($\Delta R_{SD}$) caused by the underlap versus the effective channel length ($L_{eff}$); the dotted line indicates $I_{on}$ of the abrupt $N_{SD}(y)$ device.
cell is defined by the bit-line load capacitance and the driving current of transistors in SRAM cell. Therefore, severe degradation of $R_{SD}$ can cause the read/write access failure. So, although the optimal $L_{eSD}$ should be defined to get high $V_t$, severe $I_{on}$ degradation must be avoided.

(Note from additional UFDG simulation results in Fig. 5.3 that, for all $L_{ext}$, $I_{on}$ can even be higher than that of the abrupt-$N_{SD}(y)$ counterpart (the dash line) when $L_{eSD}$ is short. Also, we see that, for the same $L_{eSD}$, $I_{on}$ is higher for longer $L_{ext}$. All of these observations can be explained by the effect of $Q_B/C_{ox}$, or the $V_t$ shift reflected in Fig. 5.2 along with the relatively small $\Delta R_{SD}$. Namely, $I_{on}$ reduction caused by the increased $R_{SD}$ in the underlap is not so significant, and thus $I_{on}$ increment caused by the $V_t$ shift via $Q_B/C_{ox}$ tends to enhance the total on-state current. Because of higher $\sigma_L$ and thus larger $V_t$ shift, this effect is more obvious for short $L_{eSD}$. These characteristics of the underlap are quite useful since the parasitic capacitance in weak inversion can still be made small by the underlap, while $I_{on}$ can be comparable to (or even larger than) that of the abrupt-$N_{SD}(y)$ case. Thus, we can utilize this interesting characteristic of the underlap to improve the device speed as we discuss in Sec. 5.5.)

For $V_t$ dependence on $w_{Si}$ variation, we confirm in Fig. 5.4 that, for thinner $w_{Si}$, $V_t$ roll-off with increasing $\sigma_L$ is a bit smaller and the absolute value of $V_t$ is much larger due to better SCE immunity. Thus, throughout our study, we assume $w_{Si} = L_g/2$, (except for $L_g < 10nm$ in the
Figure 5.4 UFDG-predicted threshold voltage ($V_t$) versus the effective channel length ($L_{eff}$) for the different fin width ($w_{Si}$); $L_{ext}$ is fixed at 12nm, and the different $\sigma_L$ defines the underlap ($L_{eSD}$) or the effective channel length ($L_{eff}$).
SRAM scaling study), rather than $L_{\text{eff}}/2$, which is the maximum fin width to control SCEs [Yan05]. Here, it should be noted that $V_t$ is defined by $\phi_c$ as in (5.1) and thus affected by $w_{\text{Si}}$ in the form of $\ln(1/w_{\text{Si}})$ [Tri05b], although $\Delta \phi_c$ for varying $w_{\text{Si}}$ is negligibly small. Also, for short $L_{\text{eSD}}$ defined by high $\sigma_L$, the contribution of $Q_B/C_{\text{ox}}$ to $\Delta V_t$ is larger for thicker $w_{\text{Si}}$ since $Q_B$ is defined by $q w_{\text{Si}} N_B$ [Tri05b]. Nevertheless, we can infer that $\Delta V_t$ caused by $\Delta w_{\text{Si}}$ is defined predominantly by SCEs through (5.2). The effects of $t_{\text{ox}}$ variation on $V_t$ should be similar to those of $w_{\text{Si}}$ variation, and thus thinner $t_{\text{ox}}$ is desirable for device optimizations. However, to control gate leakage currents without high-$k$ gate dielectric, thinning $t_{\text{ox}}$ is not a viable option for device optimizations. So, in this study, we do not consider the option of changing $t_{\text{ox}}$.

5.4 Applications on SRAM Cell

5.4.1 SRAM Cell Design

Based on $V_t$ versus $L_{\text{eff}}$ (or $L_{\text{eSD}}$) characteristics of the underlapped DG FinFETs presented in the previous subsection, we expect that, for the given $t_{\text{ox}}$ and $w_{\text{Si}}$, long $L_{\text{eSD}}$ defined by short $L_{\text{ext}}$ and low $\sigma_L$ could be a good design approach for SRAM applications. Indeed, as evident in Fig. 5.5(a), which shows UFDG-predicted read-SNM versus $L_{\text{eff}}$ (or $L_{\text{eSD}}$) of the 6-T SRAM cell shown in Fig. 5.5(b), composed of DG FinFETs with $w_{\text{Si}} = 9\text{nm}$ and $t_{\text{ox}} = 1\text{nm}$, the read-SNM generally follows the $V_t$ variations shown in Fig. 5.1. Therefore, we conclude that, for the given $t_{\text{ox}}$ and $w_{\text{Si}}$, making $L_{\text{eSD}}$ long by selecting short $L_{\text{ext}}$ and low $\sigma_L$ could be a
Figure 5.5 (a) UFDG-predicted read static noise margin (read-SNM) versus the effective channel length ($L_{\text{eff}}$) and (b) schematic of 6-T SRAM cell [Cho06]; read-SNM generally follows $V_t$-$L_{\text{eff}}$ plot in Fig. 5.1; the dotted line in (a) represents the read-SNM of the abrupt $N_{SD(y)}$ device.
good device design approach to get high $V_t$ and thus large read-SNM in SRAM applications. However, we infer that the DG FinFET design with $L_{eSD} = 3.6\, \text{nm}$, which is defined by $L_{ext} = 12\, \text{nm}$ and $\sigma_L = 6\, \text{nm}$, can be an optimal design. This is because, for very long $L_{eSD}$, there is severe reduction of $I_{on}$ as predicted in Fig. 5.3, portending possible read/write access failure. Note here that the technological lower limit of $\sigma_L$ seems to be about 5nm, and thus the defined underlap is doable. Compared to the DG FinFET design with the abrupt-$N_{SD}(y)$, this optimal DG FinFET has 29% higher $V_t$, while $I_{on}$ is only about 3% lower. It yields read-SNM equal to 209mV, 18% higher than that of the abrupt-$N_{SD}(y)$ design.

The UFDG/Spice3-predicted butterfly curve of the optimal DG FinFET design with G-S/D underlap is shown in Fig. 5.6, which also includes the butterfly curve of the abrupt-$N_{SD}(y)$ design for comparison. Higher inverter trip point and thus higher read-SNM of the device with higher $V_t$ are reflected in these butterfly curves.

Also, the benefit of high $V_t$ in the write-0 margin is shown in Fig. 5.7, even though it is very small. Therefore, by controlling SCEs via the optimal underlap, we could improve read-SNM and write-margin, as well as standby power as implied by very small $I_{off}$.

To improve the read-SNM further, high $\beta$-ratio, i.e., the size-ratio between $W/L$ of the pull-down transistor (i.e., $N_1$ or $N_2$ in Fig 5.5(b)) and $W/L$ of the access transistor (i.e., $N_3$ or $N_4$ in Fig. 5.5(b)) in the SRAM cell, is usually used. Indeed, we could increase the read-SNM to 240mV
Figure 5.6 UFDG-predicted butterfly curves of the three $L_g = 18$nm SRAM cells composed of a DG FinFET ($w_{Si} = 9$nm, $t_{ox} = 1$nm) without underlap and two underlapped DG FinFETs (via $L_{ext} = 12$nm and $\sigma_L = 6$nm) with $\beta$-ratio = 1 and 2; $V_{DD} = 1.0$V.
Figure 5.7  UFDG-predicted write-0 margin for the three different $L_g = 18\text{nm}$ SRAM cells defined in Fig. 5.6: For simulations, $V_L$ and $V_R$ in Fig. 5.5(b) are first set to 1V and 0V, respectively. Then, $V_{BL\text{-true}}$ ($V_{BL\text{-true}}$ in Fig. 5.5(b)) is swept from 1V to 0V, monitoring $V_{data\text{-true}}$ ($V_L$ in Fig. 5.5(b)).
with $\beta$-ratio = 2 via increasing the gate length of the access transistors, as indicated in Fig. 5.6. However, when high $\beta$-ratio is used, the write-margin tends to be degraded as shown in Fig. 5.7. Thus, the $\beta$-ratio should be optimized by effecting the trade-off between read-SNM and write-margin. Another approach to increase the $\beta$-ratio and thus improve read-SNM further is up-sizing the pull-down transistor over the access transistor, by using a multi-fin FinFET for the former. Our UFDG-predicted results show that this cell design approach is the most effective, enhancing the read-SNM to 269mV. However, this design might not be an optimal one due to the substantive layout-area penalty [Guo05] and write-margin reduction.

5.4.2 Sensitivity Issue in SRAM Cell

As implied in Figs. 5.1 and 5.4, thin $w_{Si}$, short $L_{ext}$, and/or low $\sigma_L$ are beneficial for better immunity to the process-induced parameter variations. To check these effects more thoroughly, the sensitivity of the optimal design developed in Sec. 5.4.1, i.e., the DG FinFET with $L_g = 18\text{nm}$, $t_{\text{ox}} = 1\text{nm}$, $w_{Si} = 9\text{nm}$, and the underlap $L_{eSD} = 3.4\text{nm}$ defined by $L_{\text{ext}} = 12\text{nm}$ and $\sigma_L = 6\text{nm}$, is examined in this subsection by varying $\sigma_L$, $w_{Si}$, and $L_g$. Even though it has been argued that the write operation is most sensitive to parametric variations [Bha05], the sensitivity study here focuses on the read-SNM fluctuations to evaluate the obtained simulation results with other ones [Cho06].
Figure 5.8 first shows UFDG/Spice3-predicted $V_t$, $I_{on}$, and read-SNM variations for the optimal DG FinFET design caused by the variation of straggle ($\Delta \sigma_L$). Here, $\Delta \sigma_L$ is assumed to be $\pm 1\text{nm}$, or $\pm 17\%$ variation from the normal $\sigma_L = 6\text{nm}$. As shown in Fig. 5.8(a), $\Delta V_t$ is less than 7%, which is much smaller than that predicted ($\Delta V_t = 24\%$ for $\Delta \sigma_L \equiv 1\text{nm}$, or 10%) by Chowdhury [Cho06]. Also, as shown in Fig. 5.8(b), $\Delta I_{on}$ due to $\Delta \sigma_L$ is less than 5%, while that predicted by Chowdhury [Cho06] is about 18% for $\Delta \sigma_L \equiv 1\text{nm}$. Because of the better $V_t$ and $I_{on}$ immunity to $\Delta \sigma_L$ of the optimally designed DG FinFET, the read-SNM variations (circles in Fig. 5.8(c)) caused by $\Delta \sigma_L$ is also very small ($< 4\%$), while the maximum $\Delta \text{SNM}$ predicted by Chowdhury [Cho06] is 13% for $\Delta \sigma_L \equiv 1\text{nm}$. The higher sensitivities in [Cho06] resulted because the DG FinFET is designed with long $\sigma_L (= 9.5\text{nm})$.

With regard to mismatch effects in the SRAM cell, we check the $\Delta \text{SNM}$ that results when an access (N4 in Fig. 5.5(b)) or pull-down transistor (N2 in Fig. 5.5(b)) is replaced with one defined by varied $\sigma_L$. As shown in Fig. 5.8(c), the predicted $\Delta \text{SNM}$ caused by the mismatch for the access transistor via $\Delta \sigma_L$ (1nm, or 17%) is small, being less than 5%. The $\Delta \sigma_L$ here defines only a small change in the $\beta$-ratio of the SRAM cell via the implied $\Delta L_{eff}$ (or $\Delta V_t$) and/or $\Delta I_{on}$ of the access transistor. As noted in the previous discussion of the $\beta$-ratio effect, a larger (2x) change in the $\beta$-ratio effects a sizable SNM variation. Similar results are predicted for a mismatched pull-down transistor via $\Delta \sigma_L$, as indicated in Fig. 5.8(c).
Figure 5.8  UFDG/Spice3-predicted sensitivity of the optimal DG FinFET design to the variation of $\sigma_L$; (a) $V_t$, (b) $I_{on}$, and (c) read-SNM variations; the mismatch effect in (c) has been checked by replacing an access/pull-down transistor with the one defined by $\Delta \sigma_L$. 
Indeed, the FinFET SRAM cell appears to be well immune to such mismatches defined by reasonable, random variations in $\sigma_L$.

Also, for $w_{Si}$ and $L_g$ variations, the optimal DG FinFET design shows good immunity with regard to $\Delta V_t$, $\Delta I_{on}$, and thus $\Delta SNM$ when compared with corresponding results predicted by Chowdhury [Cho06]. This can be explained by the well-tempered underlap along with this thin $w_{Si}$ of the optimal DG FinFET design. Table 5.1 summarizes the sensitivity of the optimal DG FinFET design to the variations of $\sigma_L$, $w_{Si}$, and $L_g$. Based on the comparison of these simulation results to those predicted by Chowdhury [Cho06], which is generally consistent with [Guo05] showing $\Delta SNM \equiv 10\%$, we confirm that the DG FinFET design, with optimal underlap, in Sec. 5.4.1 can be superior in terms of the sensitivity to the process-induced parameter variations.

Finally, it should be noted that for a hypothetical $L_g = 22$nm bulk-Si SRAM cell, $\Delta SNM$ caused by the body dopant fluctuation has been reported to be about 26% [Guo05], which is much larger than that of the 18nm DG FinFET counterpart. Also, Samsudin et al. [Sam06] reported that an $L_g = 10$nm FD/SOI single-gate (SG) device is more stable than an $L_g = 35$nm bulk MOSFET in 6-T SRAM operations. Since the reported large $\Delta SNM$ is mainly due to large $\Delta V_t$ and $\Delta I_{on}$ caused by the body dopant fluctuation, $\Delta SNM$ caused by the mismatch in the bulk SRAM cell seems to be very significant. Note that large $\Delta I_{on}$ due to large $\Delta V_t$ can cause significant SNM variations when the access transistor is in the mismatch.
Table 5.1

UFDG/Spice3-predicted sensitivity of the optimal DG FinFET design to the variations of $\sigma_L$, $w_{Si}$, and $L_g$. Note that the assumed mismatches caused asymmetric butterfly curves, and so the $\Delta SNM$ given reflects the worse-case sensitivity defined with regard to $V_L$ or $V_R$ in Fig. 5.6.

<table>
<thead>
<tr>
<th></th>
<th>$\Delta \sigma_L$</th>
<th>$\Delta w_{Si}$</th>
<th>$\Delta L_g$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(+/-) 1nm or (+/-) 17%</td>
<td>(+/-) 1nm or (+/-) 11%</td>
<td>(+/-) 2nm or (+/-) 11%</td>
</tr>
<tr>
<td>$\Delta V_t$</td>
<td>-6.3% / +3.6%</td>
<td>-6.4% / +5.2%</td>
<td>+4.1% / -6.7%</td>
</tr>
<tr>
<td>$\Delta I_{on}$</td>
<td>+5.0% / -4.9%</td>
<td>-0.3% / -0.3%</td>
<td>-0.6% / +0.1%</td>
</tr>
<tr>
<td>$\Delta SNM$</td>
<td>-3.8% / +1.3%</td>
<td>-3.9% / +2.3%</td>
<td>+0.9% / -2.8%</td>
</tr>
<tr>
<td>$\Delta SNM$ by mismatch</td>
<td>Access: -4.9% / +3.8%</td>
<td>-0.6% / +0.4%</td>
<td>+1.5% / -1.8%</td>
</tr>
<tr>
<td></td>
<td>Pull-down: +4.1% / -3.9%</td>
<td>-3.3% / +2.2%</td>
<td>-3.9% / -3.5%</td>
</tr>
</tbody>
</table>
Therefore, the nanoscale bulk SRAM cell might be limited in its usage mainly due to the access transistor mismatch. With regard to the SRAM cell size, the DG FinFET SRAM should have no area penalties because the FinFET exploits the third dimension; the layout area is defined by the half-pitch. Indeed, the layout area of the DG FinFET 6-T SRAM cell is, at the 90nm technology node [Sem05], calculated to be 0.36$\mu$m$^2$ [Guo05], which is less than that ($\approx 0.7\mu$m$^2$) of the conventional planar SRAM cell [Jun04]. Therefore, we can infer that the SRAM cell based on the nonclassical devices, including DG FinFETs, has better immunity to intrinsic parameter fluctuations than that based on bulk-Si MOSFETs, while it has no layout area penalties.

5.4.3 SRAM Cell Scaling

In the previous subsection, we have shown that the underlap in DG FinFETs is quite useful for SRAM applications since it effectively increases $V_t$, and thus improves read-SNM and write-margin. When the device is scaled down, the benefit of the underlap should be more obvious since SCE control is getting more difficult in the scaled device. So, we will now check the benefit of the underlap on the SRAM scalability by comparing read-SNM of SRAM cell composed of DG FinFETs with and without the underlap.

Fig. 5.9 shows UFDG-predicted read-SNM versus the physical gate length ($L_g$) of 6-T SRAM cell composed of DG FinFETs with and without the underlap. For the simulations, $w_{Si}$ was assumed to be $L_g/2$ to
Figure 5.9  (a) UFDG-predicted read-SNM versus the physical gate length (L_g) and (b) the read-SNM normalized by the supply voltage (V_{DD}) versus L_g plot; in (a), there are two transition points due to constant w_{Si} and V_{DD}, while there is only one transition point in (b), which is due to a constant w_{Si}.

![Graph showing UFDG-predicted read-SNM versus L_g and normalized read-SNM versus L_g.](image)
control SCE, except for the cases of $L_g < 10\text{nm}$ since $5\text{nm}$ is assumed to be the lower limit of $w_{Si}$ [Tri03a]. So, for $L_g < 10\text{nm}$, $w_{Si}$ is kept to be equal to $5\text{nm}$. For the underlap structure, we assumed a fixed S/D doping profile, i.e., $L_{ext} = 12\text{nm}$ and $\sigma_L = 6\text{nm}$, for all node applications, conservatively keeping the straggle above the noted lower limit of $\sigma_L \cong 5\text{nm}$. We would expect somewhat better results if $\sigma_L = 5\text{nm}$ were assumed, and therefore the results in Fig. 5.9 are not necessarily the best attainable with underlapped FinFETs. Finally, the supply voltage was assumed to follow the ITRS projections [Sem05].

Now, note in Fig. 5.9 that by using the underlap structure we can improve read-SNM at all nodes by more than 15%. Also, we see that the SRAM cell can be scaled down to the end of the roadmap, i.e., $L_g = 5\text{nm}$, via DG FinFETs with G-S/D underlap, while the SRAM cell via DG FinFETs without underlap has its scaling limit around at $L_g = 10\text{nm}$, mainly due to severe SCEs.

It is interesting to note that the read-SNM versus $L_g$ plot shows two transition points as marked in Fig. 5.9(a): one is at $L_g = 10\text{nm}$, where $w_{Si}$ is limited at $5\text{nm}$ and the other one is at $L_g = 7\text{nm}$, where $V_{DD}$ is limited at $0.7\text{V}$. Although it can be seen more clearly for the DG FinFET without underlap, the read-SNM decreases at a faster rate beyond the first transition point since $w_{Si} = L_g/2$ is not applicable when $L_g < 10\text{nm}$, and thus more SCEs come in. Beyond the second transition point, the reduction rate of the read-SNM is not as fast as that for $L_g > 7\text{nm}$ since
V_{DD} is not scaled. Fig. 5.9(b) shows the read-SNM normalized by V_{DD}, and thus contains only one transition point, i.e., the one at L_g = 10nm. Note that this transition point for the underlap case is more evident in Fig. 5.9(b).

Finally, in Fig. 5.10, the butterfly curve of the 6-T SRAM cell composed of L_g = 5nm DG FinFETs with the underlap defined by L_{ext} = 12nm and σ_L = 6nm is exemplified, and compared to those of the DG FinFET with β-ratio = 2 via the longer channel length for the access transistor or two fins for the pull-down transistor. As shown, we can improve read-SNM by 14% and 17% by using longer L_g for access transistor and two fins for the pull-down transistor, respectively. However, as can be seen in Fig. 5.11, the write-0 margin tends to be severely (more than 24%) degraded. Therefore, increasing the β-ratio in SRAM cell should be optimally selected via with the trade-off between read-SNM and write margin.

5.5 Device Speed Issue

5.5.1 Using Long Straggle

From the study of the relation between S/D doping profile and V_t (and thus I_{on}) in the previous subsection, we learned that I_{on} loss caused by the underlap can be minimized by using high σ_L. To quantify this effect, UFDG-predicted current-voltage characteristics of two different underlapped DG FinFET designs are shown in Fig. 5.12, along with that of the DG FinFET without underlap. As evident in the figure, the underlap
Figure 5.10 UFDG-predicted butterfly curves of the three $L_g = 5\text{nm}$ SRAM cells composed of DG FinFETs ($w_{Si} = 5\text{nm}$, $t_{ox} = 0.9\text{nm}$) with underlap (via $L_{ext} = 12\text{nm}$ and $\sigma_L = 6\text{nm}$); three SRAM cells are different from each other by the channel length for the access transistor or/and the number of fin for pull-down transistor.
Figure 5.11 UFDG-predicted write-0 margin of three different $L_g = 5$nm SRAM cell designs defined in Fig. 5.10: For simulations, $V_L$ and $V_R$ in Fig. 5.5(b) are first set to 1V and 0V, respectively. Then, $V_{BL\text{-true}}$ ($V_{BT\text{-true}}$ in Fig. 5.5(b)) is swept from 1V to 0V, monitoring $V_{data\text{-true}}$ ($V_{L}$ in Fig. 5.5(b)).
Figure 5.12 UFDG-predicted current-gate voltage characteristics (a) in log-scale and (b) linear-scale of a DG FinFET design without underlap and two underlapped designs defined by $L_{\text{ext}} = 12\text{nm}$ and $\sigma_L = 6\text{nm}$, and $L_{\text{ext}} = 18\text{nm}$ and $\sigma_L = 12\text{nm}$; $L_g = 18\text{nm}$, $t_{\text{ox}} = 1\text{nm}$, $t_{\text{si}} = 9\text{nm}$, midgap gate.
with $L_{\text{ext}} = 12\text{nm}$ and $\sigma_L = 6\text{nm}$ effectively controls SCEs and thus increases $V_t$, while it degrades $I_{\text{on}}$ some due to $\Delta R_{\text{SD}}$. For the underlap design with longer $L_{\text{ext}} (= 18\text{nm})$ and higher $\sigma_L (= 12\text{nm})$, due to small $V_t$ via the S/D dopants in the channel, i.e., the effect of $Q_B/C_{\text{ox}}$, and the small subthreshold swing $(S)$, UFDG-predicted $I_{\text{on}}$ is higher than that of abrupt-$N_{\text{SD}}(y)$ case. Here, note that the $V_t$ reduction caused by $Q_B/C_{\text{ox}}$ is ~50mV as discussed in subsection 5.3, and thus the average S/D dopant density in the channel is estimated to be $N_B \sim 1.2 \times 10^{18}/\text{cm}^3$. Since the volume of the channel is $V_{\text{ch}} = [L_{\text{eff}} (= 22.8\text{nm}) \times w_{\text{Si}} (= 9\text{nm}) \times W_{\text{eff}} (= 5w_{\text{Si}} = 45\text{nm})]$, the actual number of dopants in the channel is calculated to be $N_B \times V_{\text{ch}} = 11$. Therefore, for the underlap design with high $\sigma_L$ the random dopant effect ($\Delta N_B$) in the channel seems to be tolerable. Note, however, that if the actual dopant count in the channel were much below 10, the design would have to be refined.

Further, the subthreshold gate capacitance ($C_G$) for the two underlapped devices is much less than that of the abrupt-$N_{\text{SD}}(y)$ counterpart as clearly shown in Fig. 5.13. And, because of this reduced $C_G$, the underlapped designs show speed superiority over the abrupt-$N_{\text{SD}}(y)$ counterpart as shown in Fig. 5.14. The UFDG-predicted speed benefit of using the underlap is more than 23%. This result is consistent with that in chapter 4, which showed that the underlap is quite effective in reducing the fringe capacitance and thus the CMOS delay. Between the two underlap designs, the longer-$\sigma_L$ design shows only $\equiv 5\%$ shorter delays.
Figure 5.13 UFDG-predicted gate capacitance ($C_G$) versus gate voltage for the three 18nm DG FinFET designs defined in Fig. 5.12.
Figure 5.14 UFDG/Spice3-predicted propagation delays per stage versus supply voltage of 9-stage unloaded CMOS-inverter ring oscillators comprising three different 18nm DG FinFET designs defined in Fig. 5.12.
However, as shown in Fig. 5.15, which represents the UFDG/Spice3-predicted loaded RO delays ($\tau_{pd}$) versus load capacitance ($C_L$) of the three different 18nm DG FinFET designs defined in Fig. 5.12, the speed benefit of the longer-$\sigma_L$ design is enhanced with increasing $C_L$. This is because, when $\sigma_L$ is long to get short $L_{eSD}$, $I_{on}$ is comparable to (or even higher than) that of the abrupt-$N_{SD(y)}$ device due to large $V_t$ shift by $Q_B/C_{ox}$ in (5.1), while the parasitic capacitance in weak inversion is still small due to the defined underlap. On the other hand, the speed benefit of the short-$\sigma_L$ design over the abrupt-$N_{SD(y)}$ counterpart is degraded when $C_L$ increases, due to smaller $I_{on}$. So, using long $\sigma_L$ in the underlap design can be a good design approach to keep $I_{on}$ high and fringe capacitance small, thus improving the device speed. However, as implied in Fig. 5.1, $V_t$ of this longer-$\sigma_L$ design seems to be very sensitive to $\Delta L_{eSD}$ or $L_{ext}/\sigma_L$. Therefore, the applicability of the design with long $L_{ext}$ and $\sigma_L$ might be limited by the sensitivity issue. Thus, in the next subsection, the sensitivity issue of this optimal DG FinFET design using long $\sigma_L$ will be discussed in more detail.

5.5.2 Sensitivity to Straggle

As mentioned in the subsection 5.3, $V_t$ is defined by both SCEs and the S/D dopants in the channel, and thus very sensitive to $\sigma_L$ variations when $\sigma_L$ is long. So, in this subsection, the sensitivity issue with regard to the process induced $\Delta \sigma_L$ will be discussed for the two DG FinFET designs with the underlap defined by long and short $\sigma_L$. 
Figure 5.15 UFDG/Spice3-predicted loaded (C_L on each stage) CMOS ring-oscillator delays versus C_L, at V_{DD} = 1.0V, for the three different 18nm DG FinFET designs defined in Fig. 5.12.
Fig. 5.16 compares the $I_{on}$ and $C_G$ variations of the underlapped DG FinFET with long $\sigma_L$ to those of the counterpart with short $\sigma_L$, when the process-induced $\sigma_L$ variation is assumed to be $\pm$ 1nm. Note that $\Delta \sigma_L$ of $\pm$ 1nm corresponds to $\pm$ 8.3% and $\pm$ 16.7% variations from $\sigma_L = 12$nm and 6nm, respectively. As shown in Fig. 5.16(a), due to $\Delta \sigma_L$, $I_{on}$ of the underlapped device with longer $\sigma_L$ varies from +12% to -8%, which is much larger than those ($\pm$ 5%) of the design with shorter $\sigma_L$. On the other hand, $C_G$ variations of the two underlapped device are comparable, as shown in Fig. 5.16(b). Here, it should be noted that $\Delta C_G$ and $\Delta I_{on}$ vary in the same direction. Since the device speed correlates with $CV/I$, and $C$ and $I$ here, i.e., $C_G$ and $I_{on}$, vary in the same direction, the effects of $\Delta \sigma_L$ on the device speed should not be so significant. Indeed, for the two underlapped designs, UFDG-predicted $\Delta \tau_{pd}$ is much smaller than $\Delta I_{on}$ and $\Delta C_f$ as shown in Fig. 5.17(a). However, because of the large variations of $I_{on}$, $\Delta \tau_{pd}$ of the longer-$\sigma_L$ design reaches to $\approx$ 11% when there is heavy load; see Fig. 5.17(b). For the shorter-$\sigma_L$ design, the delay variations for the loaded case are much less than that of the longer-$\sigma_L$ design. Therefore, with regard to the device speed issue, the optimal underlap for DG FinFET devices has to be defined by the trade-off between the device performance and the sensitivity to the process-induced parameter variations.

5.6 Summary

Using Medici and UFDG simulations, we confirmed that the underlap is quite effective in controlling SCEs and thus increasing $V_t$ for
Figure 5.16 UFDG/Spice3-predicted sensitivity comparison of the two underlapped DG FinFET designs to the variation of $\sigma_L$; (a) UFDG-predicted $\Delta I_{on}$ and (b) Medici-predicted $\Delta C_G$ in weak inversion with the variation of $\sigma_L$. 

(a) $V_{DS}=50\text{mV}$

(b) $L_{ext}=18\text{nm} & \sigma_L=12\text{nm}$

- $L_{ext}=12\text{nm} & \sigma_L=6\text{nm}$
Figure 5.17 UFDG/Spice3-predicted sensitivity comparison of the two underlapped DG FinFET designs to the variation of $\sigma_L$; (a) UFDG-predicted unloaded and (b) loaded ($C_L = 0.5fF$) propagation delay variation ($\Delta\tau_{pd}$) at $V_{DD} = 1.0V$ with variation of $\sigma_L$. 
18nm FinFETs, while it sacrifices $I_{on}$ due to the increase $R_{SD}$. However, for short $L_{eSD}$ defined by long $L_{ext}$ and $\sigma_L$, $V_t$ is not defined only by SCEs but also by the S/D dopants in the channel via $Q_{B/Cox}$, and thus very sensitive to $\Delta \sigma_L$. Therefore, for SRAM applications, which require high $V_t$, the underlap has to be designed with long $L_{eSD}$, which is defined by short $L_{ext}$ and low $\sigma_L$ due to the sensitivity issue. However, because of the access time, severe $I_{on}$ reduction should be avoided, and thus the optimal underlap must be defined by the trade-off between $V_t$ and $I_{on}$. On the other hand, for the device speed issue, because of the large $V_t$ shift by $Q_B/C_{ox}$ and small $\Delta R_{SD}$ even with the underlap, $I_{on}$ of DG FinFETs with the underlap defined by high $\sigma_L$ can be comparable to (or even larger than) that of the abrupt-$N_{SD(y)}$ counterpart device, while the gate capacitance in weak inversion is reduced by the underlap. Thus, with the benefit in $I_{on}$ and $C_G$ afforded by the long $\sigma_L$, we could enhance the device speed benefit in the underlapped devices, especially with heavy loads. However, because of the relatively severe sensitivity to $\Delta \sigma_L$, the underlap using higher $\sigma_L$ should be carefully designed.
CHAPTER 6
SUMMARY AND SUGGESTIONS FOR FUTURE WORK

6.1 Summary

This dissertation addressed performance projections, design optimization, and physical modeling issues of nonclassical nanoscale CMOS, including fully depleted SOI single-gate, double-gate, and triple-gate MOSFETs. The major contributions of the research are summarized as follows.

In chapter 2, we projected device characteristics and CMOS performances of nonclassical UTB CMOS technologies optimized at the $L_g = 28\text{nm}$ node, and compared them with that of classical, hypothetical bulk-Si CMOS at this node. With the same UTB thickness, the DG devices were shown to be far superior to the FD/SG devices with regard to SCE control, and generally superior to SG devices, including bulk-Si devices, with regard to speed because of higher drive currents. However, an interesting insight was noted. For light loads and moderate supply voltages, a suboptimal FD/SG design (with the same $t_{Si}$) for both LOP and HP applications was found to yield speeds comparable to the DG designs, even though its current drives are much lower and its SCEs are much more severe. This surprising comparison was shown to be a result of the FD/SG devices having much lower intrinsic gate capacitance, which is due to their thick BOX and higher subthreshold swing, and hence deferred onset of
significant inversion-charge capacitance. At lower $V_{DD}$, however, the DG designs are much faster because of their much higher drive currents. When the FD/SG CMOS design was optimized by aggressive scaling of the UTB thickness, its high-$V_{DD}$ speed diminished (but was still comparable to that of DG CMOS) because of higher gate capacitance at intermediate gate voltages, while its low-$V_{DD}$ speed improved due to increased current. Compared to the nonclassical CMOS, the predicted delay of the bulk-Si/SG CMOS was much longer due to its high gate capacitance in the weak/moderate-inversion region, in addition to the areal source/drain junction capacitance, and relatively low drive current limited by polysilicon-gate depletion.

In chapter 3, three-dimensional numerical simulations of DG and TG FinFETs having undoped thin bodies revealed the significance of bulk-inversion current in $I_{on}$, as well as $I_{off}$, and the consequent insignificance of the commonly defined effective gate width in comparisons of DG and TG currents. In fact, we inferred that the proper $W_{eff}$ for DG FinFETs is $h_{Si}$, which correlates with the total (surface plus bulk) inversion charge; whereas a meaningful $W_{eff}$ cannot be directly defined for TG FinFETs. The new insights revealed herein explain why the DG FinFET provides nearly the same $I_{on}$ as the TG counterpart for fin aspect ratios as small as two, but especially for higher $R_f$ which is desirable and doable. Due to the relatively small increase in $I_{on}$ of TG FinFETs, over the DG counterparts with moderate $R_f$, the advantage of TG devices in gate layout-area efficiency is not significant. The insights thus
further solidify our notion, based initially on $W_{\text{eff}}$-implied TG layout-area inefficiency [Yan05] (and on the fact that a TG FinFET, with a thin top dielectric and moderate $R_f$, is more difficult to fabricate than a DG FinFET [Mat05]), that the third top gate is neither desirable nor beneficial.

In chapter 4, we showed, with 2-D numerical device simulations, that the parasitic fringe capacitances in nonclassical nanoscale MOSFETs, e.g., DG FinFETs, are very significant, with important $V_{GS}$ dependences due to the gate-source/drain underlap that in fact reduces the capacitance. With physical insights from the device simulations, we developed an analytical model for the parasitic capacitance, including inner- and outer-fringe components, and a BOX-fringe component for FD/SOI MOSFETs, all with dependences on $V_{GS}$ and on the underlap structure. The model was verified generally by the numerical simulations, and implemented in our process/physics-based compact model (UFDG-3.5). With UFDG in Spice3, we showed, via ring-oscillator simulations, that reducing the parasitic capacitance via optimal underlap design can be quite effective in improving nanoscale DG CMOS speed, which is basically defined by a trade-off regarding the capacitance and source/drain series resistance. Nonetheless, we stress the severity of the G-S/D fringe-capacitance effect on speed shown by our simulations. We found that this parasitic capacitance is predominant in limiting nanoscale DG CMOS speed, even when moderated by an optimal G-S/D underlap.

In chapter 5, using Medici/UFDG simulations, we confirmed that the G-S/D underlap is quite effective in controlling SCEs and thus
increasing $V_t$ for a given DG FinFET structure. However, for short $L_{eSD}$ defined by long $L_{ext}$ and high $\sigma_L$, $V_t$ is not defined only by SCEs but also by the S/D dopants in the channel via $Q_B/C_{ox}$. Therefore, for SRAM applications, which requires high $V_t$, the underlap has to be designed with long $L_{eSD}$, which is defined by short $L_{ext}$ and low $\sigma_L$. However, because of the access time, severe $I_{on}$ reduction should be avoided, and thus the optimal underlap must be defined by the trade-off between $V_t$ and $I_{on}$. On the other hand, for the device speed issue, we can use high $\sigma_L$ to get short $L_{eSD}$. Because of the large $V_t$ shift by $Q_B/C_{ox}$ and small $\Delta R_{SD}$ even with the underlap, on-state current can be comparable to that of abrupt-$N_{SD}(y)$ FinFET, while the parasitic capacitance for weak inversion is made smaller by the underlap. Thus, with the benefit in $I_{on}$ and $C_G$ afforded by high $\sigma_L$, we could enhance the device speed benefit in the underlaped devices, especially with the heavy loads. However, because of the relatively severe sensitivity to $\Delta \sigma_L$, the underlap using higher $\sigma_L$ should be carefully designed.

6.2 Suggestions for Future Work

In chapter 3, we showed that there is significance bulk-inversion current in $I_{on}$, as well as $I_{off}$, for DG and TG FinFETs with undoped UTBs. Because of this bulk inversion, there are several interesting features in the nonclassical devices with undoped bodies. For example, as introduced in chapter 3, the proper $W_{eff}$ for DG FinFETs is $h_{Si}$, which correlates with the total (surface plus bulk) inversion charge. Also, because of the smaller
transverse electric field in the undoped channel, the carrier mobility in the channel tends to be higher than that of the conventional doped bulk-Si devices. Finally, the bulk inversion reduces the gate capacitance, and thus $C_G$ of DG FinFETs tends to reach $C_{ox}$ at higher gate voltages. On the other hand, with regard to the abnormal $C_G$ reduction in the saturation region, which is observed in Medici [Med04] simulation, the bulk inversion can be a possible reason as explained in appendix B. However, Taurus [Tau04] does not predict this abnormality of $C_G$. Therefore, it is worth examining the difference between Medici and Taurus, that causes the different predictions for the saturation $C_G$, and verify if the $C_G$ reduction is a real effect or not. Also, if the $C_G$ reduction is real, appendix B should be re-evaluated as a possible explanation for that.

In the underlap study, we have assumed that the source/drain series resistance ($R_{SD}$) is not a function of $V_{DS}$. This is true when $L_{eSD}$ is short. However, when $L_{eSD}$ is long, $R_{SD}$ is severely affected by the drain voltage. Since, for SRAM applications, the pull-down transistor and the access transistor are working in different region of operation, this $V_{DS}$-dependence of $R_{SD}$ is quite important. For example, when the zero-state voltage is defined in SRAM, the pull-down transistor is operating in the linear region, while the access transistor is in the saturation region. So, without the proper analytical model, we can not predict the exact zero-state voltage and thus read-SNM. Therefore, the $V_{DS}$-dependence of $R_{SD}$ should be studied to show the range of $L_{eSD}$ which can be used without the
concern about this $V_{DS}$-dependence, and, if necessary, a $V_{DS}$-dependent $R_{SD}$ model should be developed.

Also, as shown in chapter 5, the S/D dopants in the channel tend to reduce $V_t$ of nonclassical undoped-UTB devices when $\sigma_L$ is high to get short $L_{eSD}$. This effect was not important in the SRAM applications since, for high $V_t$ and read-SNM, long $L_{eSD}$ via low $\sigma_L$ is beneficial. However, with regard to the device speed issue, high $\sigma_L$ seems to yield higher $I_{on}$, which is basically enabled by the large $V_t$ shift caused by the S/D dopants in the channel. The main problem here is that UFDG does not include any physical parameter or model to account for this $V_t$ shift. Note that the $V_t$ versus $\sigma_L$ study, via UFDG, in chapter 5 has been done by modulating the work function of the gate electrode to properly account for the $V_t$ shift by the S/D dopants in the channel, with the aid of Medici. Therefore, with regard to the S/D dopants in the channel, a proper physical parameter or model is required for UFDG to predict $V_t$ characteristics accordingly.

Even though the access time is critical in defining if the SRAM read or the write operation is successfully carried out or not, a comprehensive study has not been done in this dissertation mainly due to the uncertainty of the bit-line capacitance. In chapter 5, the read or the write operation failure has been avoided by defining the optimal underlap, which does not reduce $I_{on}$ much, with the assumption that $I_{on}$ of the abrupt-$N_{SD}(y)$ device is large enough to make the SRAM cell operate properly. However, because of the difficulty in controlling the gate leakage
current without the high-k gate dielectric material and the increased S/D series resistance due to the extremely thin SOI film, the \( I_{on} \) requirement tends to be more difficult to meet with the device scaling. Thus, the access-time simulation, with the proper value of the bit-line capacitance, seems to become more important, and thus should be done to complement the SRAM cell study.
APPENDIX A
UPGRADES/REFINEMENTS OF UFDG CHARGE MODEL

A.1 Modeling of Junction Depletion Charge

The depletion region charge in the body-source (/drain) junction, which defines the reverse-bias junction capacitance, is defined as

\[ Q_{S, \text{dep}} = -qW_{\text{eff}}N_{\text{Body}}t_{\text{Si}}w_{\text{dep}}, \tag{A.1} \]

where, \( N_{\text{Body}} \) and \( t_{\text{Si}} \) depend on the structure of the device. The depletion width at the neutral body/source junction, \( w_{\text{dep}} \), is defined as

\[ w_{\text{dep}} = w_{\text{dep}0} \left[ 1 - \frac{V_{BS, \text{eff}}}{V_{\text{bi}}} \right], \tag{A.2} \]

where

\[ w_{\text{dep}0} = \sqrt{\frac{2\varepsilon_{\text{Si}}V_{\text{bi}}}{qN_{\text{Body}}}}, \tag{A.3} \]

\[ V_{BS, \text{eff}} = V_{\text{bi}} - V_{t} \ln \left[ 1 + \exp \left( \frac{V_{\text{bi}} - V_{BS}}{V_{t}} \right) \right], \tag{A.4} \]

and

\[ V_{\text{bi}} = V_{\text{bi}0} - (\text{SCEB})V_{t} \ln \frac{N_{S}}{n_{b}}. \tag{A.5} \]

\( N_{S} \) is the doping density in the source and \( n_{b} \) is the carrier density in the extension, where gate bias dependent modulation is negligible, and fixed
at $10^{19}$ cm$^{-3}$. SCEB is a user defined model parameter in UFDG and can be a positive value between 0 and 1 [Cho06]. The built-in potential $V_{bi0}$ in (A.5) is assumed to be

$$V_{bi0} = \frac{E_g}{2} + V_t \ln \frac{N_{Body}}{n_i}, \quad (A.6)$$

where $n_i$ is the intrinsic carrier density at a given temperature. Similar equations for the depletion charge at the body/drain junction, $Q_{D,dep}$, require an effective $V_{BD}$, i.e., $V_{BD,eff}$.

$Q_{S,dep}$ and $Q_{D,dep}$ are added to the total body charge in UFDG, and $-Q_{S,dep}$ and $-Q_{D,dep}$ are added to the total source and drain charge, respectively, for charge neutrality. Therefore, we have

$$Q_S = Q_S - Q_{S,dep}, \quad (A.7)$$

$$Q_D = Q_D - Q_{D,dep}, \quad (A.8)$$

and

$$Q_{Body} = Q_{Body} + Q_{S,dep} + Q_{D,dep}. \quad (A.9)$$

Here, it should be noted that the modeled depletion charge, $Q_{S/ D,dep}$, at the junction between the source/drain and the body is a few electron charges, which are about two order magnitude less than the fringe capacitance charge [Kim06]. Thus, we can neglect this junction depletion charge in terms of its amount. However, this model seems to be necessary in UFDG for the model stability. Therefore, even though $Q_{S/ D,dep}$ does not contribute to $Q_{Body}$, this model is implemented into UFDG.
A.2 Upgrading of Electron Charge Model in Weak Inversion

The diffusion length, $L_e$, in the channel is defined as [Yeh05]

$$L_e = L_{\text{eff}} - L_s - L_d,$$

(A.10)

with

$$L_s = \frac{2[V_{bi} - \Psi_{\text{min}}(x)]}{\left| \frac{\partial \Psi(x, y)}{\partial y} \right|_{y = 0}},$$

(A.11)

and

$$L_d = \frac{2[V_{bi} - \Psi_{\text{min}}(x) + V_{DS}]}{\left| \frac{\partial \Psi(x, y)}{\partial y} \right|_{y = L_{\text{eff}}}},$$

(A.12)

where $L_{\text{eff}} = L_g + L_{es} + L_{ed}$ in the weak inversion region and $\Psi_{\text{min}}(x)$ is the minimum potential in the body. In (A.11) and (A.12), it is assumed that $\Psi$ in the body near the source and drain is a linear function in $y$ direction as shown in Fig. A.1. The built-in potential, $V_{bi}$, at the virtual source-body boundary is expressed in (A.5). Between $y = L_s/2$ and $y = L_s/2 + L_e$ for a given $x$, we assume a constant $\Psi$, i.e., $\Psi_{\text{min}}$. This assumption tends to make the potential and, consequently, the total charge ($Q_n$) in the body underestimated. However, the predicted error is negligible since $Q_n \propto e^{(\Psi)}$.

The total charge, $Q_n$, in the body can be obtained by three partitioned channel charges in $y$, i.e., $Q_{nB}$, $Q_{nS}$, and $Q_{nD}$ as defined in Fig. A.1. Each charge is composed of four partitioned channel charge in $x$, assuming linear variations of potential in $x$ between $\Psi_{\text{min},(xj)}$ and $\Psi_{\text{min},(x_{j+1})}$ with $j = 0, 1, 2, \text{ and } 3$. The partitioned channel charge in the
Figure A.1 Representative potential variations in \( y \) at a given \( x \) and corresponding linear approximations.
middle of the body, $Q_{nBj}$ with $j=0$, 1, 2, and 3, is given by

$$Q_{nBj} \equiv \frac{-qW_{eff}^2 \Gamma_{ni}}{4N_{Body}} \frac{V_t}{\Psi_{min(xj)} - \Psi_{min(xj+1)}} \left[ e^{\left(\frac{\Psi_{min(xj+1)} - \Psi_{min(xj)}}{V_t}\right)} - e^{\left(\frac{\Psi_{min(xj+1)} - \Psi_{min(xj+2)}}{V_t}\right)} \right]. \quad (A.13)$$

And, the partitioned charge, $Q_{nSj}$, in the body near the source is expressed by

$$Q_{nSj} = \frac{-qW_{eff}^2 \Gamma_{ni}}{4N_{Body}} \int_0^{L_s/2} e^{\frac{\Psi_S(y)}{L_s/2}} \frac{V_t}{\Psi_S(y)} dy \quad (A.14)$$

$$= \frac{-qW_{eff}^2 \Gamma_{ni}}{8N_{Body}} \left[ e^{\frac{\Psi_S(y=L_s/2)}{V_t}} - e^{\frac{\Psi_S(y=0)}{V_t}} \right]. \quad (A.15)$$

with $\Psi_S(y=0) = V_{bi}$ and $\Psi_S(y=L_s/2) = (\Psi_{min(xj)} + \Psi_{min(xj+1)})/2; j=0, 1, 2$ and 3.

We get the similar expression with regard to $Q_{nDj}$, the partitioned channel charge in the body near the drain, by replacing $L_s/2$ and $\Psi_S(y=0) = V_{bi}$ with $L_d/2$ and $\Psi_S(y=L_{eff}) = V_{bi} + V_{DS}$, respectively. The total body charge, $Q_n$, is then expressed by

$$Q_n = \sum_j Q_{nBj} + \sum_j Q_{nSj} + \sum_j Q_{nDj}, \quad (A.16)$$

where $j=0, 1, 2, \text{ and } 3$. And, finally, the front gate total charge is expressed as

$$Q_{Gf} \rightarrow Q_{Gf} - Q_n, \quad (A.17)$$

with

$$Q_{Gf} = C_{oxf} W_{eff} L_e (V_{GfS} - \Psi_{min(x0)} - \Phi_{MS}), \quad (A.18)$$

where, $\Phi_{MS}$ is the work-function difference between the front gate metal.
and the body. Then, the back gate total charge can be represented by the charge neutrality, i.e.,

\[ Q_{Gb} = -(Q_s + Q_d + Q_{Gf} + Q_b + Q_{ff} + Q_{fb}), \]  
(A.19)

with

\[ Q_s = \frac{\left(\sum Q_{nBj}\right)}{2} + \sum Q_{nSj}, \]  
(A.20)

\[ Q_d = \frac{\left(\sum Q_{nBj}\right)}{2} + \sum Q_{nDj}, \]  
(A.21)

and

\[ Q_b = -(W^L_{eff}qN_{Body}t_{Si}), \]  
(A.22)

where \( Q_{ff} \) and \( Q_{fb} \) are the fixed front- and back-oxide charges.

However, due to the existence of the fringe capacitance charge [Kim06], which is about one order magnitude larger than the electron charge in weak-inversion region, this model does not seem to be necessary in UFDG. So this model is not implemented in UFDG.
APPENDIX B
DG MOSFET GATE CAPACITANCE IN SATURATION REGION

Medici-predicted $C_G$-$V_{GS}$ characteristics shown in Fig. B.1 provide an interesting, novel feature of symmetrical double-gate (SDG) MOSFETs with undoped UTBs, i.e., unusual gate capacitance ($C_G$) reduction at saturation region. Because of the channel length modulation, $C_G$ is smaller at saturation than that at linear region as represented in Fig. B.1. With increasing $V_{GS}$ at high $V_{DS}$, the channel length modulation is less effective and thus $C_G$ should increase with increasing $V_{GS}$ as it does for low $V_{DS}$. However, as shown in Fig. B.1, $C_G$ for strong inversion at high $V_{DS}$ is decreasing with increasing $V_{GS}$.

However, Taurus [Tau04] simulations do not show this abnormality, and hence the abnormal $C_G$ reduction in saturation region may or may not be real. Therefore, more study, with a simulator that more physically models the carrier transport in the channel (which governs $C_G$ at high $V_{DS}$), is needed. In this appendix, we, believing the Medici simulation result, suggest a possible explanation for this unusual $C_G$ reduction in saturation region.

One possible explanation is based on $V_{GS}$-dependent inner-fringe capacitance ($C_{if}$), which is negated at strong inversion by the inversion charge screening. At high $V_{DS}$, due to the less inverted charges
Figure B.1 Medici-predicted gate capacitance at low and high $V_{DS}$ versus gate voltage for the $L_g=18\,\text{nm}$ DG nMOSFET ($t_{Si}=9\,\text{nm}$, $t_{ox}=1\,\text{nm}$, midgap gate) without G-S/D underlap and outer fringe capacitance, i.e., $C_{of}=0$ by $t_g=0$. 
near the drain, $C_{if}$ might not be completely screened out at saturation region. With increasing $V_{GS}$, inversion electrons increase and thus $C_{if}$ can be effectively screened out by the inversion-charge screening. Consequently, $C_{if}$ (and thus $C_G$) can decrease with increasing $V_{GS}$. However, Medici-predicted inverted-electron density of SDG devices in Fig. B.1 is more than $\sim 10^{18}/\text{cm}^3$ near the drain, which seems to be large enough to screen out $C_{if}$. In addition, the observed $C_G$ reduction is more severe for the device with thinner $t_{Si}$, which has smaller $C_{if}$ [Kim06], and even larger than the total $C_{if}$ for the device with $t_{Si}=3\text{nm}$ as shown in Fig. B.2. Therefore, we can infer that $C_{if}$ is not the reason for the negative $C_G$ slope at saturation region observed in Fig. B.1.

Another possible reason for the unusual $C_G$ reduction at saturation region is DICE, which is 2-D effect and thus loses its effect, i.e., charge enhancement, with increasing $V_{GS}$. To check this effect, a relatively long-channel device ($L_g = 0.1\mu\text{m}$) with very thin $t_{ox}(=1\text{nm})$ and $t_{Si}(=10\text{nm})$, in which 2-D effect can be ignored, is simulated with Medici for $C_G-V_{GS}$ characteristics. In Fig. B.3, the saturation-$C_G$ of this device is compared to those of other devices with thicker $t_{Si}(=50\text{nm})$ or higher body doping ($5\times 10^{18}/\text{cm}^3$). We see for this device that $C_G$ is still decreasing at saturation region even though DICE or 2-D effect is virtually eliminated with extremely thin $t_{ox}$ and $t_{Si}$. Rather, $C_G$ reduction could be negated for the device with thicker $t_{Si}$ and completely removed by high body doping as
Figure B.2 Medici-predicted saturation gate capacitance versus gate voltage for the L\textsubscript{g}=18nm DG nMOSFET (t\textsubscript{ox}=1nm and midgap gate) with various SOI film thickness; DG devices are assumed to have abrupt N\textsubscript{SD}(y) without the G-S/D underlap.
Figure B.3 Medici-predicted saturation gate capacitance versus gate voltage for three different $L_g=0.1 \mu m$ DG MOSFETs with $t_{ox}=1nm$; devices are separated from each other by the SOI film thickness and/or the body doping.
shown in Fig. B.3. These results imply that the saturation-\( C_G \) reduction in Fig. B.1 is not caused by DICE.

The real explanation can be induced from Fig. B.4, which shows the electron-density profile across the SOI film for three different devices in Fig. B.3. From Figs. B.3 and B.4, it is obvious that \( C_G \) reduction is more severe for the device with higher inversion charges in the bulk and can be negated by eliminating the bulk inversion charges. Thus, we can infer that the negative \( C_G \) slope observed in Fig. B.1 is fundamentally based on the bulk-inversion [Kim05b], which is a strong function of \( t_{Si} \) and the body doping.

Based on this conclusion, we can explain the unusual \( C_G \) reduction at saturation region as follows. For the relatively low \( V_{GS} \) at high \( V_{DS} \), the bulk inversion charge \( (Q_{inv(Bulk)}) \) is dominant, while the surface inversion charge \( (Q_{inv(Sur)}) \) will be more dominant with increasing \( V_{GS} \). When \( Q_{inv(Sur)} \) is getting larger, the increasing rate of \( Q_{inv(Bulk)} \), i.e., \( C_{inv(Bulk)} \), should be getting smaller due to the surface charge screening. Then, the increasing rate of the total inversion charge \( (Q_{inv}) \), i.e., \( C_{inv} \), will decrease since \( Q_{inv} \) is defined by the sum of \( Q_{inv(Sur)} \) and \( Q_{inv(Bulk)} \). Consequently, based on the definition of the capacitance, i.e., \( C = dQ/dV \), the inversion capacitance \( (C_{inv} = dQ_{inv}/dV_{GS}) \) can have a negative slope at saturation region as exemplified in Fig. B.5. Note here that \( C_G \) reduction is observed only for high \( V_{DS} \). This is because bulk inversion can be a
Figure B.4 Medici-predicted electron density profile across the SOI film for three different devices in Fig. B.3.

- $t_{Si}=10\text{nm},$ Undoped
- $t_{Si}=50\text{nm},$ Undoped
- $t_{Si}=50\text{nm}, N_{\text{body}}=5\times10^{18}\text{cm}^{-3}$
Figure B.5 Integrated inversion charge ($Q_{inv}$) at $y=\frac{L_g}{2}$ and $V_{DS}=1.0V$, and its differentiation ($dQ_{inv}/dV_{GS}$), which is reflecting $C_{inv}$, of the DG device in Fig. B.1, i.e, $L_g=18$nm DG nMOSFET with $t_{ox}=1$nm and $t_{Si}=9$nm.
predominant effect when $V_{DS}$ is high enough to lower the surface potential and increase the potential in the bulk.

To support our explanation that $C_{inv}$ decreases due to the decreasing $C_{inv(Bulk)}$, the inverted electron density ($n$) and the differentiation of it, i.e., $dn/dV_{GS}$, at the surface ($x = 0$ or $t_{Si}$) and the bulk ($x = t_{Si}/2$) are plotted in Fig. B.6 for the device in Fig. B.5. Note here that the total inversion charge $Q_{inv}$ is defined by $qnt_{Si}$ and thus $C_{inv}$ is proportional to $dn/dV_{GS}$. Then, $C_{inv(Sur)}$ and $C_{inv(Bulk)}$ can be reflected by $dn(x=0 \text{ or } t_{Si})/dV_{GS}$ and $dn(x = t_{Si}/2)/dV_{GS}$, respectively. In Fig. B.6, we clearly see that $dn(x = t_{Si}/2)/dV_{GS}$ has a negative slope beyond about $V_{GS} = 0.5V$. Thus, so does $C_{inv(Bulk)}$ even though $V_{GS}$-dependence of the surface inversion layer thickness is not accounted here. Consequently, we confirm that the increasing rate of bulk inversion charge is decreasing with increasing $V_{GS}$ due to the inversion charge screening. Then, since $C_G$ for SDG devices can be expressed by [Kim01]

$$
C_G = \frac{2C_{ox}}{C_{ox} + \frac{1}{C_{inv}}} \tag{B.1}
$$

with $C_{inv} = C_{inv(Sur)} + C_{inv(Bulk)}$. $C_G$ can have the negative slope shown in Fig. B.1, following $C_{inv(Bulk)}$ reflected by $dn(x = t_{Si}/2)/dV_{GS}$. Here, $C_{inv(Sur)}$ and $C_{inv(Bulk)}$ are assumed to be in parallel as shown in Fig. B.7 because both the surface and the bulk inversion-layer are connected to the source (or ground).
Figure B.6 Medici-predicted surface and bulk electron density (n) at $y=L_g/2$ and $V_{DS}=1.0V$, and its differentiation (dn/dV$_{GS}$) of the DG device of Fig. B.1; $C_{inv(Sur)}$ and $C_{inv(Bulk)}$ is reflected by dn(x=0 or $t_{Si}$)/dV$_{GS}$ and dn(x=$t_{Si}/2$)/dV$_{GS}$, respectively.
Figure B.7 Equivalent circuits of a DG MOS capacitor in the strong inversion region; $C_{\text{inv(Sur)}}$ and $C_{\text{inv(Bulk)}}$ are in parallel here because both the surface and the bulk inversion-layer are connected to the source (or ground).
REFERENCE LIST


BIOGRAPHICAL SKETCH

Seung-Hwan Kim was born in Seoul, Korea. He received the B.S. and M.S. degrees in metallurgy and materials science from Hanyang University, Ansan, Gyeonggi-Do, Korea, in 1995 and 1997, respectively. In 2002, he received the M.S. degree in electrical and computer engineering from the University of Florida, Gainesville. Since 2002, he has been pursuing a Ph.D. degree in electrical engineering at the University of Florida, Gainesville.

In the summer of 2002 and 2004, he was an intern at the Process Development Team and the TCAD Team, respectively, of Samsung Electronics, Korea. In 2005, he was selected as one of the recipients of the Korean Graduate Student Research Award from the University of Florida’s Korean student scholarship and loan fund. His research involves simulation, analyses, design, and physical modeling of non-classical nanoscale CMOS, such as fully depleted SOI single- and multi-gate devices.