Title
Variability and Heterogeneous Integration of Emerging Device Technologies

Author
Leung, Gregory

Publication Date
2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Variability and Heterogeneous Integration of Emerging Device Technologies

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Electrical Engineering

by

Gregory Kwong-Wah Leung

2015

ABSTRACT OF THE DISSERTATION

Variability and Heterogeneous Integration

of Emerging Device Technologies

by

Gregory Kwong-Wah Leung

Doctor of Philosophy in Electrical Engineering

University of California, Los Angeles, 2015

Professor Chi On Chui, Chair

The continued push for microelectronics scaling has driven many changes in modern transistor design, such as the adoption of non-planar, multi-gate architectures (e.g., FinFETs) starting at the 22nm node. It is envisioned that other solutions such as junctionless FETs (JL-FETs), tunnel FETs (TFETs), or heterogeneous materials integration may be needed to sustain the pace of Moore's law beyond 14nm. To assess the viability of these emerging devices prior to commercial investment, we must consider the impact of process variations such as line edge roughness (LER) and random dopant fluctuation (RDF), both of which are major concerns in the nanoscale regime. The challenges associated with dimensional scaling also compel us to explore heterogeneous integration as a possible end-of-roadmap solution for future micro- and nanoelectronics.

In this dissertation, we first present our findings on the impact of LER and RDF variability on FinFETs, JL-FETs, and TFETs targeted for sub-32nm generations. Using technology computer-aided design (TCAD) simulations combined with physical descriptions by which LER and RDF affect the intrinsic operation of different FETs, we compare the impact of LER and RDF on the emerging candidates of interest. We extend the study to include III-V FETs as well to determine if materials like InGaAs are inherently more or less affected by variability compared to equivalently designed silicon devices. Second, we study how heterogeneous integration (HGI) of different material systems can drive a new approach toward improving circuit and system performance outside of traditional scaling concepts. To this end, we develop a cross-layer evaluation framework (spanning process, device, and circuit-level perspectives) to assess the potential benefits of InGaAs/Ge-based HGI circuits against silicon-only technology. To give credence to the framework, we also present experimental work in developing a nanotransfer printing process to enable feature-level HGI in real-world settings. Third, we present a method to fabricate coplanar supercapacitors onto silicon substrates for integration with microelectronic circuits. Along with experimental demonstrations, we also develop a physical TCAD model to enable simulation-based design exploration and optimization of on-chip supercapacitors for integrated circuit applications.

Ultimately, the insights gained in this study will help guide the semiconductor industry to choose next-generation device technologies which are best suited for commercial adoption with process variability and the potential for heterogeneous integration in mind.

The dissertation of Gregory Kwong-Wah Leung is approved.

Puneet Gupta

Bruce Dunn

Subramanian Iyer

Chi On Chui, Committee Chair

University of California, Los Angeles

2015

TABLE OF CONTENTS

Chapter 1 *Introduction* 1

Chapter 2 *Inversion-Mode Silicon FinFET Variability* 5

2.1 Background 5
2.2 IM-FinFET Modeling 7
2.3 Line Edge Roughness Modeling 9
2.4 LER-Induced Variability in IM-FinFETs 12
2.5 Random Dopant Fluctuation Modeling 15
2.6 RDF-Induced Variability in IM-FinFETs 18
2.7 Circuit-Level Variability Impact 20
2.8 Summary 22

Chapter 3 *Junctionless Silicon FET Variability* 24

3.1 Background 24
3.2 JL-FET Modeling 25
3.3 LER-Induced Variability in JL-FETs 28
3.4 RDF-Induced Variability in JL-FETs 32
3.5 Circuit-Level Variability Impact 37
3.5.1 Variability Impact on 6T SRAM Cells 40
3.5.2 LER Impact on Logic Circuit Variability 44
3.6 Summary 48

Chapter 4 *Silicon Tunnel FET Variability* 50

4.1 Background 50
4.2 TFET Modeling 51
4.3 Baseline TFET Scaling 55
4.4 LER-Induced TFET Variability 59
4.5 RDF-Induced TFET Variability 62
4.6 Summary 65

Chapter 5 *Interactions between LER and RDF in Nonplanar FET Variability* 68

5.1 Background 68
5.2 Modeling Approach 69
5.3 IM-FinFET Joint Variability 72
5.4 JL-FinFET Joint Variability 76

5.5 TFET Joint Variability 79
5.6 Summary 82
5.7 Appendix: Mean Parameter Shifts 84

Chapter 6 *Silicon vs. III-V Junctionless FET Variability* 85

6.1 Background 85
6.2 Effects of Degenerate Carrier Screening 88
6.3 Baseline Design & Performance of Si and InGaAs JL-FETs 91
6.4 Modeling RDF: Approach and Limitations 97
6.5 RDF in Doped Semiconductor Slabs 99
6.6 RDF in Silicon and InGaAs JL-FETs 102
6.7 Summary 109
6.8 Appendix I: Details on NEGF Simulations and Scattering Mechanisms 110
6.8.1 Impurity Scattering 110
6.8.2 Phonon and Surface Roughness Scattering 113
6.9 Appendix II: Effect of Barrier Height on Subthreshold Swing 114

Chapter 7 *Heterogeneous Integration Technology* 117

7.1 Background 117
7.2 Nanotransfer HGI Process: Proposed Concept 122
7.3 Nanotransfer HGI Process: Experimental Work 126
7.3.1 Previous Work: Integration of GaAs NR Arrays with Si on Si/SiO_2 126
7.3.2 Experiment #1: Transfer of High Aspect Ratio GaAs NR Arrays to Si/SiO_2 128
7.3.3 Experiment #2: Transfer of High Aspect Ratio InAs NR Arrays to Si/SiO_2 135
7.4 HGI Evaluation Framework 140
7.4.1 Objective of the Framework 140
7.4.2 Device Modeling 142
7.4.3 Alignment Error and Transfer Accuracy 143
7.4.4 Transfer Yield and Performance Loss Considerations 145
7.4.5 HGI Impact on Circuit Layout and Design Rules 146
7.4.6 Projected HGI Benefits 152
7.4.7 Evaluation Summary 162
7.5 Cost Analysis 162
7.6 Summary 166
7.7 Appendix I: Experimental Procedure for GaAs Transfer to SiO_2 168

7.8 Appendix II: Experimental Procedure for InAs Transfer to SiO_2 169

Chapter 8 *Supercapacitors for Microelectronics* 170

8.1 Background 170
8.2 Process Flow Template 172
8.3 Experimental Procedure 174
8.3.1 Fluoroalkylsilane (FAS) Treatment of Exposed SiO_2: 174
8.3.2 Selective Carbon Deposition by Self-Assembly: 175
8.3.3 Ionogel Synthesis: 176
8.3.4 Electrochemical Characterization: 176
8.4 Experimental Results and Discussion 177
8.4.1 Millimeter-Scale Gold-Ionogel EDLCs 177
8.4.2 Sub-Millimeter-Scale Carbon-Ionogel EDLCs 179
8.4.3 Micrometer-Scale Well EDLCs 181
8.4.4 Benchmarking 183
8.5 EDLC Simulation and Modeling 185
8.5.1 Simulation Details 187
8.5.2 Coplanar EDLC Modeling 191
8.6 Summary 195
8.7 Appendix I: Formulas for EDLC Circuit Elements 197
8.8 Appendix II: Explanation for Series Resistance in Fig. 83(b) 197

Chapter 9 *Conclusion* 200

References 205

LIST OF FIGURES

Fig. 1. Schematic of the FinFET structure. A vertical fin made up of silicon is straddled by a metal gate layer running perpendicular to the fin length. The metallurgical gate length L_g and the fin body thickness T_{fin} are indicated in the diagram. The effective channel width is the sum of T_{fin} and $2 \times H_{fin}$, where H_{fin} is the fin height. If the gate oxide covering the top surface of the fin is much larger than on the sidewalls, then the structure resembles a simple double-gate MOSFET and the channel width can be approximated as simply twice the fin height. 5

Fig. 2. Schematic of the 2D structure used to model IM-FinFET devices (32nm case shown). The structure represents a planar cut across the fin height and parallel to the plane of the wafer. 7

Fig. 3. Gaussian LER patterns corresponding to σ_{LER} = 1 nm and λ = 5, 15, and 50 nm. Short segments of these patterns were used as inputs to the simulated FinFETs. 10

Fig. 4. Simulated IM-FinFET structures with and without 1 nm LER along the fin sidewalls. ... 11

Fig. 5. Resist (left) and spacer (right) IM-FinFET device variability as a function of LER amplitude and technology node. Markers indicate actual simulated data while solid lines indicate best fits. Note the zoomed scale for spacer IM-FinFET data compared to resist IM-FinFET data. 12

Fig. 6. Quadratic rise in average arc length for spacer FinFETs due to LER as a function of root-mean-square amplitude. The nominal arc length corresponds to a 13 nm channel length for the data shown. 14

Fig. 7. Effective doping profiles resulting from RDF in 32, 21, and 15nm IM-FinFET devices. The effective channel length becomes nonuniform and reduces on average at smaller nodes. 17

Fig. 8. RDF-induced variability in IM-FinFETs as a function of fin height and technology node. 19

Fig. 9. Overall flow of the circuit benchmark evaluation process. 20

Fig. 10. Schematic comparison of (a) junctionless and (b) inversion-mode FETs and their associated doping profiles. From [35]. 24

Fig. 11. Electron mobility plots in 32nm IM- and JL-FinFETs at $V_{GS} = V_{DS} = V_{DD} = 0.9$ V. The channel mobility is consistently higher in IM-FinFETs compared to JL-FinFETs due to reduced impurity and surface roughness scattering at these geometries. 27

Fig. 12. Representative 32nm JL-FinFET with 1 nm LER applied to the fin edges. 28

Fig. 13. Representative 32nm JL-FinFETs with and without RDF applied. The effective doping concentrations in both cases are shown with the same color legend. 28

Fig. 14. Resist JL-FinFET device variability as a function of LER amplitude and technology node. Markers indicate actual simulated data while solid lines indicate best fits..... 29

Fig. 15. Electron density plots for two representative 32nm JL-FinFETs showing the inadvertent formation of a conducting channel due to fin LER at V_{GS} = 0.1 V, and V_{DS} = 0. White lines indicate depletion region boundaries..... 30

Fig. 16. RDF-induced variability in JL-FinFETs as a function of fin height and technology node. 32

Fig. 17. (a) Relative variation and (b) absolute variation of JL-FinFET performance due to L_g scaling from 22 nm to 13 nm with H_{fin} = 10 nm. 33

Fig. 18. (a) Relative variation and (b) absolute variation of JL-FinFET performance due to T_{fin} scaling from 9.6 to 6.4 nm with H_{fin} = 10 nm. 34

Fig. 19. Electron density plots in a representative 32nm JL-FinFET (H_{fin} = 10 nm) with and without RDF, showing the inadvertent formation of a conducting channel in the off state due to a surplus of dopants in the channel for the device with RDF. 37

Fig. 20. Overview of the variability evaluation framework. The evaluation of 6T SRAM cells (left) and microprocessor circuits (right) are divided into two vertical branches as illustrated..... 38

Fig. 21. Matching of baseline FinFET (a) transfer and (b) output curves between TCAD simulation and compact modeling. 40

Fig. 22. Comparison of σI_{on} and $\sigma V_{T,sat}$ extracted from 200 samples between TCAD simulations and fitted variability models for (a) JL FinFETs and (b) IM FinFETs show a good fit. .. 40

Fig. 23. Nominal SNM as a function of working V_{cc} for high density design JL FinFET 6T SRAM cells. Note that for successive technology nodes, SNM and $V_{cc,min}$ decrease when the other is held fixed..... 41

Fig. 24. V_{ccmin} as a function of technology node and LER amplitude for JL and IM FinFET 6T SRAM. The SNM constraint is 100 mV with 99% yield in the left panel, and 10 mV with 99.9% yield in the right panel. 43

Fig. 25. (a) Nominal clock period and clock period increase (mean shift and variation) and (b) nominal leakage power and leakage power increase (mean shift and variation) due to LER variation (σ_{LER} = 0.6 nm) for IM and JL-FinFET-based MIPS processors at typical clock speeds. 46

Fig. 26. (a) Increase in clock period mean and (b) variation of critical clock period as a function of technology node and LER amplitude for JL- & IM-FinFET circuit benchmark (Cortex-M0)..... 47

Fig. 27. (a) Increase in leakage power mean and (b) variation of leakage power as a function of technology node and LER amplitude for JL- & IM-FinFET circuit benchmarks (Cortex-M0). 47

Fig. 28. (a) Structure of an *n*-type silicon TFET with *p*-type source, intrinsic channel, and *n*-type drain. (b) Band diagrams in the "off" state for all-silicon TFET and a silicon TFET with a Ge source. (c) Respective band diagrams in the "on" state. From [61]. 50

Fig. 29. Simulated *n*-type silicon DG TFET structure along with the doping strategy used in this work. 51

Fig. 30. Examples of simulated 20/5 TFETs with and without LER and RDF. 54

Fig. 31. (a) Raw $I_D - V_{GS}$ curves for the ideal TFETs with V_{GS} swept from 0 to V_{DD} and $V_{DS} = V_{DD}$. (b) Nominal TFET performance versus body thickness scaling from 3 nm to 12 nm in terms of the metrics V_T, I_{on}, I_{off}, and SS. Solid markers indicate the performance of 20/5 and 20/10 TFETs. Curves for the 1-D Schrodinger model are incomplete due to convergence problems. 55

Fig. 32. On-state ($V_{GS} = V_{DS} = V_{DD}$) current density maps for the ideal 20/5 TFET (left) and 20/10 TFET (right), along with the energy band diagrams along two horizontal cut lines: one along the body midsection (solid) and another near the silicon-SiO_2 interface (dashed). In thin body TFETs, the energy bands are sufficiently lowered by the high gate voltage to induce BTBT along the midsection in addition to the two surface channels. In thick body TFETs, significant BTBT only occurs along the two surface channels and not along the midsection. 56

Fig. 33. Off-state ($V_{GS} = 0$ and $V_{DS} = V_{DD}$) current density maps for the ideal 20/5 TFET (left) and 20/10 TFET (right), along with the energy band diagrams along two horizontal cut lines: one along the body midsection (solid) and another near the silicon-SiO_2 interface (dashed). In thin body TFETs, the barrier height for trap-assisted tunneling is lowered by the close proximity of the gate to all vertical locations in the channel compared to thick body TFETs where the gate loses control of the midsection, resulting in a larger barrier to prevent significant tunneling through traps. 57

Fig. 34. Device-level variability of 20/5 and 20/10 TFETs due to body LER with σ_{LER} ranging from 0 to 1 nm. Markers indicate actual data while lines indicate best fits. 59

Fig. 35. Effect of a specific LER pattern ($\sigma_{LER} = 1$ nm) on the drive current of a 20/5 TFET compared to an equivalent IM-FinFET. Only the doping and work function are different between the two, all other parameters are identical. 61

Fig. 36. Device-level variability of 20/5 and 20/10 TFETs due to RDF for different device heights *H* from 10 to 40 nm. 63

Fig. 37. (a) On-state and (b) off-state band diagrams for 20/5 TFETs with and without RDF along the channel surface and channel midsection. RDF causes the source-channel tunneling path to slightly widen in the on-state, while the direct source-to-drain trap-

assisted tunneling path shortens in the off-state. (c) Corresponding $I_D - V_{GS}$ curves with and without RDF, showing severe degradation in performance predicted from the RDF model..... 64

Fig. 38. Examples of simulated structures containing body/fin LER and RDF: (left) 32nm IM-FinFET, (center) 32nm JL-FinFET, and (right) 20/10 TFET. All devices are shown with a height of 20 nm. 71

Fig. 39. (a) Comparison of expected IM-FinFET variability when LER and RDF are assumed uncorrelated versus direct simulations with LER and RDF present. (b) Percentage error incurred when assuming independence of LER and RDF compared to actual simulated values. 73

Fig. 40. (a) Comparison of expected JL-FinFET variability when LER and RDF are assumed uncorrelated versus direct simulations with LER and RDF present. (b) Percentage error incurred when assuming independence of LER and RDF compared to actual simulated values. 77

Fig. 41. Simulated resistors with and without LER & RDF. L = 40 nm, W = 5 nm, and H = 10 nm in the structures shown with σ_{LER} = 1 nm and nominal doping of 2×10^{19} cm^{-3}. 78

Fig. 42. (a) Comparison of expected TFET variability when LER and RDF are assumed uncorrelated versus direct simulations with LER and RDF present. (b) Percentage error incurred when assuming independence of LER and RDF compared to actual simulated values. 80

Fig. 43. Distributions of $V_{T,sat}$ and I_{on} for 15nm IM and JL-FinFETs and 20/5 TFETs with LER and RDF. The IM-FinFETs and TFETs have noticeable skew while JL-FinFETs appear normal symmetric. 81

Fig. 44. Highest room temperature mobility of electrons (red) and holes (blue) versus semiconductor lattice constant in inversion layers and quantum wells. Data points which lie along a drawn arrow indicate different amounts of semiconductor biaxial strain and their respective strain-enhanced mobility. From [72]. 85

Fig. 45. Electron injection velocities of InGaAs and InAs HEMTs and Si MOSFETs as a function of gate length. The saturation of InGaAs channel mobility at shorter gate lengths indicates near-ballistic operation; this observation is supported by ballistic Monte Carlo simulations which lie coincident with the experimental data. From [72]. 86

Fig. 46. Effective conduction band and valence band density of states in various semiconductors. Data taken from [91]. 88

Fig. 47. Conduction band DOS in (a) $In_{0.53}Ga_{0.47}As$ and (b) Si quantum wells calculated using 2-D atomistic tight-binding (TB) and effective mass (EM) Hamiltonians, compared with the equivalent 3-D DOS normalized by the well thickness. 89

Fig. 48. Comparison of (a) n as function of Fermi energy E_F, (b) dn/dE_F versus n, and (c) screening length versus n in InGaAs and silicon. 2-D calculations are performed using the tight binding DOS. 2-D values of n and dn/dE_F are normalized to 3-D by dividing by the channel thickness T = 6.4 nm. 90

Fig. 49. Device characteristics for (a) InGaAs and (b) silicon n-type JL-FETs with 2×10^{19} cm^{-3} channel doping from ballistic (solid lines) and scattering (dashed lines) NEGF simulations. For InGaAs, the curve including SR scattering is calculating assuming a roughness amplitude Δ = 1.76 nm. 92

Fig. 50. Spectral current along center of InGaAs and Si 2×10^{19} cm^{-3} doped devices in the off- (V_{GS} = 0) and on-states (V_{GS} = 0.73 V). The green lines indicate the position of the source Fermi energy and white lines mark the first subband edge. Note the different energy scales for InGaAs and silicon. 94

Fig. 51. Nominal I_D-V_G curves for 15nm (a) Si and (b) InGaAs JL-FETs showing TCAD calibrations performed against NEGF simulations. The upper curves in each panel correspond to the log scale on the left while the lower curves correspond to the linear scale on the right. 96

Fig. 52. Examples of JL-FETs exhibiting RDF generated from the Sano method. For reference, the nominal structure (without RDF) is shown in the upper left panel, having a uniform doping concentration of 2×10^{19} cm^{-3} 97

Fig. 53. Comparison of spatial fluctuations along a z-cutline in (a) electrostatic potential and (b) electron density in 100×100×100 nm^3 n-Si and n-InGaAs resistor slabs resulting from RDF. The nominal doping concentration (without RDF) for both slabs is 10^{20} cm^{-3}. Both slabs have exactly the same number and spatial arrangement of dopants. 100

Fig. 54. Average fluctuations of (a) potential, (b) electron density, (c) current, and (d) normalized current in ensembles of 20 × 20 × 20 nm^3 Si and InGaAs slabs with RDF for different nominal doping concentration N. The ensemble size is 100 slabs for each combination of material and N. The applied voltage is 10 mV in (c) and (d). In (c), the Si curves are scaled by 5× for visual clarity. In (d), the current fluctuations are normalized to the ideal current values when RDF is absent from the slab. 101

Fig. 55. Comparison of raw n-type and p-type InGaAs and Si JL-FET variability due to RDF for the metrics $V_{T,lin}$, $V_{T,sat}$, I_{on}, I_{off}, SS, and DIBL. 102

Fig. 56. Comparison of normalized n-type and p-type InGaAs and Si JL-FET variability due to RDF for the metrics $V_{T,lin}$, $V_{T,sat}$, I_{on}, I_{off}, SS, and DIBL. The standard deviations are normalized to the baseline values given in Table 17. 103

Fig. 57. Nominal conduction band diagrams along the center of the channel in n-type (a) InGaAs and (b) Si JL-FETs under the following bias conditions (displayed from top to bottom): off-state, saturation threshold, linear threshold, and on-state. The electron quasi-Fermi energy level is shown in dashed lines for each bias condition. The inset in (a) compares

the bands at linear and saturation threshold near the top of the barrier, indicating greater degeneracy at $V_G = V_{T,lin}$ compared to $V_{T,sat}$. 104

Fig. 58. (a) Dependence of SS and (b) DIBL on the nominal channel doping in Si and InGaAs JL-FETs. The sensitivity of SS and DIBL to *N* is lower for *n*-InGaAs JL-FETs due to degeneracy effects. 106

Fig. 59. Pseudo 2-D model subthreshold swing as a function of barrier height using geometric and material parameters from Table 17. The sensitivity of SS to barrier height (equal to $qdSSdETOB$) is also shown at $ETOB$ = 0.2 eV and $ETOB$ = 0.4 eV for Si and InGaAs, respectively. 116

Fig. 60. Process flow sequence for NTP-based HGI. 122

Fig. 61. Possible wafer-scalable concept of III-V/Ge HGI on Si realized through a repeatable "step and transfer" NTP process. 125

Fig. 62. (a) HGI demonstration of 400 nm wide GaAs and Si nanoribbon arrays formed by NTP on SiO_2/Si substrate with mm^2 area coverage. (b) Measured overlay error (16 μm) after aligned transfer and source/drain electrode formation using optical lithography. 127

Fig. 63. MBE-grown layer stack for *n*-GaAs/$Al_{0.8}Ga_{0.2}As$/GaAs substrate. 128

Fig. 64. CD-SEM image of as-etched GaAs NRs in citric acid/hydrogen peroxide solution. The nominal width of each ribbon is 0.5 μm, whereas the actual measured width is 0.781 μm. 129

Fig. 65. Optical micrographs of GaAs NR arrays being undercut by selective etching of the underlying $Al_{0.8}Ga_{0.2}As$ after (a) 0, (b) 6, and (c) 9 min in dilute BOE solution. White areas correspond to the top layer *n*-GaAs while violet corresponds to the bottom GaAs layer. After a 9 min undercut, some NRs began collapsing as indicated by translucent segments at random locations. 130

Fig. 66. A set of 10^{18} cm^{-3} *n*-doped GaAs nanoribbon (L/W/T = 400/0.75/0.03 μm) arrays transferred to SiO_2/Si. Each individual array is nominally composed of ten parallel ribbons. Discontinuities along the nanoribbons indicate broken segments resulting in <100% yield. 132

Fig. 67. Optical micrographs of a completely processed L_G = 7.5 μm *n*-GaAs JL-FET on SiO_2 substrate. 134

Fig. 68. MBE-grown layer stack for *n*-InAs/$Al_{0.4}Ga_{0.6}Sb$/GaSb substrate. 136

Fig. 69. Time progression of InAs NR undercutting by AlGaSb etching after (a) 7.5 min, (b), 8.0 min, and (c) 8.5 min in dilute NH_4OH solution. Each 200×500 μm^2 rectangular area contains an array of 500 parallel NRs. The dark blue regions correspond to NRs that are insufficiently undercut, while the lavender regions correspond to bent and/or collapsed NRs that received an excessive undercut. The etch front illustrated by the dashed orange

outline in (c) corresponds to NR portions that are on the verge of collapse. In (d), broken portions of InAs from the etch front were successfully transferred to SiO_2, but nothing else. 137

Fig. 70. Optical micrographs of a fully processed *n*-InAs JL-FETs on SiO_2 after annealing in N_2 for (a) 30 min at 350°C and (b) 60 min at 450°C. Discoloration of the InAs near the metal lines is visible after annealing at 450°C. 138

Fig. 71. Two-terminal I-V measurements performed on InAs NRs on SiO_2. The electrode separation is 5 μm and the InAs thickness is 15 nm for each device. The effective width of each device is unknown but lies somewhere between 0.5 and 2 μm. 139

Fig. 72. Power-delay tradeoff for 15nm InGaAs/Ge and Si/Si built Cortex-M0 generated by PROCEED [152]. 141

Fig. 73. NEGF (symbols) and model fit (lines) $|I_D|$-$|V_{GS}|$ curves for Si, Ge, and InGaAs double-gate FinFETs. The dashed line represents the NEGF Si NFET simulation using nanoMOS [153]. All simulations are with drain bias V_{DS} = 0.73 V. Inset: double-gate structure used for simulations. 143

Fig. 74. Schematic layouts for heterogeneous FinFET inverters from NTP without fin trimming. The area of transfer uncertainty indicates the region where PFET fins can land due to misalignment. 147

Fig. 75. (a) Schematic layout for a row of heterogeneous FinFET inverters made with NTP and fin trimming. (b) The effect of transfer misalignment with fin trimming is now absent within each cell except at the buffer areas on ends of a row. 150

Fig. 76. (a) Probability of successful fin placement as a function of transfer misalignment and allotted overlay margin. (b) Alignment yield versus average cell area in reduced MIPS processor. (c) Optimal OLM value search to maximize alignment yield per cell area. . 153

Fig. 77. Delay versus area for 15nm InGaAs/Ge (HGI) and Si/Si (non-HGI) inverters for different σ and CHs (a) without fin trimming and (b) with fin trimming. The inset is a magnified view of the dashed region in (b). 154

Fig. 78. Protocol for block-level HGI design. A grid of dummy filling cells (red cells) are inserted pre-placement to represent the effect of finite fin length, and standard cells (blue cells) are then placed in between the filling cells. 157

Fig. 79. Post-synthesis (pre-P&R) normalized delay and area of (a) MIPS and (b) AES designs. Post-P&R normalized delay and power of MIPS and AES designs with MAFL of (c,d) 5 μm and (e,f) 1 μm, respectively. In each panel the reported data is normalized to the largest observed delay, power, or area values as indicated by the data labels. The design rules (i.e., OLM values) are chosen to ensure 95% yield in all cases. 159

Fig. 80. Total interconnect length as a function of maximum allowed fin length for HGI-based (a) MIPS and (b) AES designs. 161

Fig. 81. Estimated cost breakdown to implement 22nm FinFET technology in different process scenarios. For HGI processes, the integrated material pair is realized by either nanotransfer printing (NTP) or nanoheteroepitaxy (NHE). 163

Fig. 82. Illustration of a generic process flow for integrating planar carbon-ionogel EDLCs on a silicon substrate. The numbers 1-7 indicate the sequence of processing steps. 173

Fig. 83. Measured (a) capacitance and (b) series resistance dispersions for B30 gold-ionogel EDLCs. Solid (dashed) lines correspond to the 1 μm (3 μm) gel devices. The inset in (a) is an optical micrograph of one of the measured B33 devices. The color fringing in the ionogel is indicative of film thickness variations. 177

Fig. 84. Comparison of series capacitance and resistance dispersions for 10 μm gap supercapacitors with and without KB in neat ionic liquid. 180

Fig. 85. (top) Optical microscope image of KB self-assembled electrodes coated with 100 μm thick ionogel. (bottom) SEM images of ketjen black particles self-assembled on the gold electrodes. 180

Fig. 86. Comparison of areal capacitance and series resistance measurements on the 10 μm gap carbon-ionogel supercapacitor. 181

Fig. 87. Capacitance and series resistance dispersions for 15×15 μm well gold-ionogel supercapacitors. The inset shows optical micrographs of two such devices, denoted “A” and “B”. The well openings are indicated by the black square outlines. 182

Fig. 88. Simulation model for the sandwich configuration EDLC. 188

Fig. 89. Plots of (a) electrostatic potential and (b) space charge in a simulated sandwich EDLC for different DC voltage biases. The supercapacitor has a 4 nm thick electrolyte, 2.8 Å thick Stern layer, and 1 nm thick electrodes. 189

Fig. 90. Equivalent circuit diagram for the simulated EDLC model showing individual contributions from the Stern, diffuse, and bulk electrolyte regions. 190

Fig. 91. Simulation model for the coplanar configuration EDLC. 191

Fig. 92. Comparison of experimental planar EDLC capacitance and series resistance versus simulated values from the TCAD setup. 192

Fig. 93. Effect of scaling the (a) electrode length, (b) gap distance, and (c) electrolyte thickness on the capacitance dispersion of planar EDLCs. The electrolyte is assumed to be [BMIM][BF4] ionic liquid with 1 mS/cm conductivity and 2.8 Å Stern layer. 193

Fig. 94. Spatial current distributions in coplanar EDLCs with 1 um and 10 um thick electrolyte films with a 100 mV applied voltage. Crowding effects are more visible in thinner electrolytes. 194

Fig. 95. Comparison of experimental planar EDLC capacitance and series resistance versus simulated values from the TCAD setup (a)–(b) with the addition of a fixed 150 pF capacitance to the simulated results for consistency with the experimental measurements. In (c)–(d), the extra 150 pF capacitor is removed, demonstrating the series resistance drop at f > 3 kHz is introduced by the Solartron. 198

LIST OF TABLES

Table 1. Nominal Parameters for Simulated IM-FinFETs 8

Table 2. Delay and Leakage Mean and Sigma over All Benchmarks with 1 nm LER for Resist (R) and Spacer (S) FinFET Technologies 22

Table 3. Nominal Parameters for Simulated JL-FinFETs 26

Table 4 Allowed Tuning Range of Fitted Compact Model Parameters 39

Table 5. Nominal SNM and SNM Loss from Variability for JL-FinFET Technologies 44

Table 6. Circuit Benchmarks 45

Table 7. Mean Shift and Standard Deviation of Timing and Leakage for Six Benchmark Circuits 48

Table 8. Nominal Parameters for Simulated TFETs 52

Table 9. Comparison of Average Versus Nominal TFET Performance With and Without RDF 63

Table 10. Nominal Parameters for Simulated FETs 70

Table 11. Inversion-Mode FinFET Variability from LER and RDF 72

Table 12. Suppressed LER-RDF Interactions in 15nm IM-FinFETs with L_g = 50 nm 75

Table 13. Junctionless FinFET Variability from LER and RDF 76

Table 14. Comparison of Resistor Current Variability from LER and RDF 79

Table 15. TFET Variability from LER and RDF 79

Table 16. Mean Parameter Shifts Relative to Baseline Values 84

Table 17. Nominal Parameters for Silicon and InGaAs JL-FETs 91

Table 18. Simulated I_{on} (in mA/μm) for 2×10^{19} cm^{-3} JL-FETs with Different Scattering Models. 93

Table 19. Calibrated TCAD Parameters 95

Table 20. Differences in InGaAs JL-FET Variability Based on Carrier Model 107

Table 21. Differences in *n*-JL-FET Variability Based on Calibration Setting 108

Table 22. Doping-Dependent Self-Energy Parameters 112

Table 23. Modified 15nm Design Rules for Different Process Scenarios 153

Table 24. Area, Delay, and Power of HGI Standard Cells Compared to Non-HGI Cells. 156

Table 25. Condensed Process Sequences for HGI and Non-HGI Options 164

Table 26. Process Flow for GaAs NR Transfer to SiO_2 168

Table 27. Process Flow for InAs NR Transfer to SiO_2 169

Table 28. Specified Dimensions of "B30" Ionogel on Bare Gold Supercapacitors. 177

Table 29. Benchmark Comparison against Other On-Chip Supercapacitors from Recent Literature. 183

ACKNOWLEDGEMENTS

I would first like to express my deepest appreciation and gratitude to my faculty advisor Professor Chi On Chui for all the guidance, inspiration, and the many opportunities he has provided to me throughout my career as a graduate student. His invaluable mentoring played a key role in my development as an academic scholar and researcher, and I am forever indebted to his time, dedication, and commitment to my graduate studies over the past six years.

Next, I would like to thank all of my friends and colleagues who I have worked with at UCLA, many of whom inspired me and helped me with my research throughout the years. Sincere thanks to Andrew Pan for sharing his endless knowledge in all things physics and the Spurs, Kaveh Shoorideh for his help in setting up device simulations and showing me how to characterize electronic gizmos, Yufei Mao for helping me with experiments and showing me the ropes around the Nanolab, Hyung-Suk Yu for sharing his knowledge in memory, Shaodi Wang for all his help and work in our device-circuit collaborations, Kyeong-Sik Shin, Kun-Huan Shih, and Jorge Kina for lending their experience and expertise in the transfer printing experiments, Leland Smith and Jon Lau for their tireless dedication in getting new supercapacitors made week after week, and our lab newcomers Wuran Gao, Ablai Akhazhanov, Raghav Gupta, Rowan Fakhro, and Xin Li for accepting the torch.

I would also like to thank the other members of my doctoral committee, Professors Puneet Gupta, Bruce Dunn, and Subramanian Iyer for their support in collaborative projects and helping me reach the culmination of my journey towards the Ph.D.

Last but certainly not least, I would like to thank my family and friends for their unconditional love and support, not just during my college career but throughout my life to this point and for many more years to come. Mom and Dad, you made everything possible and I can never fully

repay the sacrifices you made for me to get where I am today. I owe my accomplishments to you and hope you will be proud of this significant achievement and all to follow! Ken and Tara, thank you for all the love and being there for me throughout the years. Thank you to my buddies Rob, Aaron, and Melvin for the good times playing and talking video games, tech stuff, and well, just everything really! And finally, thank you to my furry friends Mocha and Java for always watching over from above. I love and miss you both.

PREFACE

Chapters 2 and 3 contain reproduced material from the following published works:

> G. Leung and C. O. Chui, "Variability in nanoscale FinFET technologies," *Toward Quantum FinFET* (edited by W. Han and Z. M. Wang), Springer Publishing, 2013.
>
> S. Wang, G. Leung, A. Pan, and C. O. Chui, "Evaluation of digital circuit-level variability in inversion-mode and junctionless FinFET technologies," *IEEE Trans. Electron Devices*, vol. 60, pp. 2186-2193, 2013.
>
> G. Leung, L. Lai, P. Gupta, and C. O. Chui, "Device and circuit level variability caused by line edge roughness for sub-32nm FinFET technologies" *IEEE Trans. Electron Devices*, vol. 59, pp. 2057-2063, 2012.
>
> G. Leung and C. O. Chui, "Variability impact of random dopant fluctuation on nanoscale junctionless FinFETs" *IEEE Electron Device Lett.*, vol. 33, pp. 767-769, 2012.
>
> G. Leung and C. O. Chui, "Variability of inversion-mode and junctionless FinFETs due to line edge roughness" *IEEE Electron Device Lett.*, vol. 32, pp. 1489-1491, 2011.

Chapter 4 contains reproduced material from the following published work:

> G. Leung and C. O. Chui, "Stochastic variability in silicon double-gate lateral tunnel field-effect transistors," *IEEE Trans. Electron Devices*, vol. 60, pp. 84-91, 2013.

Chapter 5 contains reproduced material from the following published work:

> G. Leung and C. O. Chui, "Interactions between line edge roughness and random dopant fluctuation in nonplanar field-effect transistor variability," *IEEE Trans. Electron Devices*, vol. 60, pp. 3277-3284, 2013.

Chapter 6 contains reproduced material from the following published works:

> A. Pan, G. Leung and C. O. Chui, "Junctionless silicon and $In_{0.53}Ga_{0.47}As$ transistors—Part I: Nominal device evaluation with quantum simulations," *IEEE Trans. Electron Devices*, early access, 2015.
>
> G. Leung, A. Pan, and C. O. Chui, "Junctionless silicon and $In_{0.53}Ga_{0.47}As$ transistors—Part II: Device variability from random dopant fluctuation," *IEEE Trans. Electron Devices*, early access, 2015.

Chapter 7 contains material from the following article currently under review:

> G. Leung, S. Wang, A. Pan, P. Gupta, and C. O. Chui, "An evaluation framework for nanotransfer printing based feature-level heterogeneous integration in VLSI circuits," *IEEE Trans. VLSI Systems*, under review.

Chapter 8 contains material from the following article currently under review:

> G. Leung, L. Smith, J. Lau, B. Dunn, and C. O. Chui, "Carbon-ionogel supercapacitors for integrated microelectronics," *IOP Nanotech*., under review.

VITA

Education

09/2010 – 08/2015 UNIVERSITY OF CALIFORNIA, LOS ANGELES, California, USA
Ph.D. candidate in Electrical Engineering

08/2008 – 06/2010 UNIVERSITY OF CALIFORNIA, LOS ANGELES, California, USA
M.S. in Electrical Engineering

08/2004 – 05/2008 UNIVERSITY OF CALIFORNIA, BERKELEY, California, USA
B.S. in Electrical Engineering and Computer Science / Materials Science and Engineering

Publications

G. Leung, L. Smith, J. Lau, B. Dunn, and C. O. Chui, "Carbon-ionogel supercapacitors for integrated microelectronics," *IOP Nanotech.*, under review.

G. Leung, S. Wang, A. Pan, P. Gupta, and C. O. Chui, "An evaluation framework for nanotransfer printing based feature-level heterogeneous integration in VLSI circuits," *IEEE Trans. VLSI Systems*, under review.

A. Pan, G. Leung and C. O. Chui, "Junctionless silicon and $In_{0.53}Ga_{0.47}As$ transistors—Part I: Nominal device evaluation with quantum simulations," *IEEE Trans. Electron Devices*, early access, 2015.

G. Leung, A. Pan, and C. O. Chui, "Junctionless silicon and $In_{0.53}Ga_{0.47}As$ transistors—Part II: Device variability from random dopant fluctuation," *IEEE Trans. Electron Devices*, early access, 2015.

G. Leung and C. O. Chui, "Variability in nanoscale FinFET technologies," *Toward Quantum FinFET* (edited by W. Han and Z. M. Wang), Springer Publishing, 2013.

G. Leung and C. O. Chui, "Interactions between line edge roughness and random dopant fluctuation in nonplanar field-effect transistor variability," *IEEE Trans. Electron Devices*, vol. 60, pp. 3277-3284, 2013.

S. Wang, G. Leung, A. Pan, and C. O. Chui, "Evaluation of digital circuit-level variability in inversion-mode and junctionless FinFET technologies," *IEEE Trans. Electron Devices*, vol. 60, pp. 2186-2193, 2013.

G. Leung and C. O. Chui, "Stochastic variability in silicon double-gate lateral tunnel field-effect transistors," *IEEE Trans. Electron Devices*, vol. 60, pp. 84-91, 2013.

G. Leung, L. Lai, P. Gupta, and C. O. Chui, “Device and circuit level variability caused by line edge roughness for sub-32nm FinFET technologies” *IEEE Trans. Electron Devices,* vol. 59, pp. 2057-2063, 2012.

G. Leung and C. O. Chui, “Variability impact of random dopant fluctuation on nanoscale junctionless FinFETs” *IEEE Electron Device Lett.*, vol. 33, pp. 767-769, 2012.

G. Leung and C. O. Chui, “Variability of inversion-mode and junctionless FinFETs due to line edge roughness” *IEEE Electron Device Lett.*, vol. 32, pp. 1489-1491, 2011.

G. Leung and C. O. Chui, “Impact of line edge roughness and device scaling on double-gate FinFET variability,” in *4th IEEE Int. Workshop on DFM&Y,* 2010.

G. Leung, “Impact of line edge roughness on sub-32nm FinFETs,” M.S. thesis, Dept. Elec. Eng., Los Angeles, CA, 2010.

Chapter 1

Introduction

The exponential growth of the semiconductor industry over the past four decades [1] has been made possible in large part by the continued scaling of microelectronic devices from 10μm feature sizes in 1970 to 14nm in 2015. Improvements in chip performance, density, and cost per function have been realized through advancements in modern field-effect transistor (FET) design and state-of-the-art manufacturing technologies. The past decade, in particular, has witnessed some of the most ambitious changes in FET design in attempt to combat increasing concerns over large standby power dissipation and fundamental speed limitations. These include the adoption of high-k gate stacks [2], ultra-thin body (UTB) or silicon-on-insulator (SOI) technologies [3], and strained silicon channels [4] targeting sub-0.1μm generations. Most recently, the commercial transition to 3-D multi-gate (MG) FETs for the 22nm node in 2011 [5] was seen as a momentous step forward beyond the planar technologies which had been in place since Gordon Moore first coined the ubiquitous industry driver known as Moore's law.

Despite major progress over the years in extending the lifetime of silicon-based complementary metal-oxide-semiconductor (CMOS) technology, consensus holds that radical innovations will be needed to continue scaling into the nanometer regime. Many novel technologies have been proposed with the intention of improving one or more aspects of FET design, including the structural configuration or the operational control mechanism. Improvements in the former generally aim to enhance the electrostatic control of the gate over the channel region through UTB or MG configurations, and may involve the use of quasi 1-D nanowires (NWs), carbon nanotubes (CNTs), or 2-D sheet materials such as graphene for the semiconducting channel. Improvements in the latter generally aim to circumvent some physical limitation in traditional CMOS devices,

such as the kT/q subthreshold swing limit, the difficulty in forming highly abrupt source/drain junctions, or limited carrier velocities in silicon channels. As an example, proposed solutions to these issues are tunnel FETs (TFETs), junctionless FETs (JL-FETs), and heterogeneous integration of Group IV/III-V FETs respectively. There is currently an abundance of new and active research to investigate the performance and fabrication of devices using these novel technologies both from academia and industry.

With each new generation, however, challenges associated with process variations must be faced which shape the manufacturing and design aspects of integrating new technologies. The impact of these variations inevitably become more and more significant as device scales shrink, making variability a major concern for integrated circuit (IC) scaling to the nanometer regime. Example sources of variability in transistors include: line edge or width roughness (LER/LWR), random dopant fluctuation (RDF), oxide thickness variation (OTV), work function variation (WFV), and many others. When presented in actual devices, these forms of variability manifest themselves via fluctuations in the performance of individual (and otherwise identically designed) devices, resulting in unpredictable performance and behavior. Designers have faced these problems for many years in planar CMOS technology and yet, despite the wealth of knowledge surrounding variability effects in planar CMOS, the issue continues to get worse for future generations.

Even more concerning is the relative uncertainty faced by today's semiconductor industry in deciding which technology solutions to invest in for future commercial adoption. Despite the rapid increase of published literature in recent years covering incremental gains in the performance of post-CMOS technologies (e.g. JL-FETs, TFETs, etc.), there is a distinct lack of understanding

whether or not certain device technologies should merit serious consideration from a manufacturability standpoint, especially considering the implications of process variability found in modern foundry tools. For example, does the absence of source/drain junctions in JL-FETs, alone, imply that such transistors are more manufacturable than traditional inversion-mode (IM) FETs, or are JL-FETs actually more vulnerable to process variability such that the benefits of their purported manufacturing ease are negated in the end? Are TFETs, by nature of their tunneling operation, more or less affected by the same variations compared to thermally operated IM-FETs or JL-FETs? Is there a point where the heterogeneous integration of Ge *p*-type and III-V *n*-type FETs becomes advantageous over Si CMOS in terms of performance versus layout area when restrictions imposed by manufacturing and variability are considered? If so, how and when should we make that transition? Lastly, are there ways in which we can harness the power of nonstandard electronic devices such as electrochemical supercapacitors for conventional on-chip microelectronic applications by way of heterogeneous integration? With these questions in mind, a framework which enables some form of early assessment/evaluation of performance, variability, and potential for heterogeneity in emerging device technologies would be an invaluable asset to the semiconductor industry moving forward.

In the following chapters, we seek to answer the aforementioned questions and others related to them. In Chapter 2, we present our methodology to evaluate the variability impact of LER and RDF on IM-FinFETs targeted to meet sub-32nm generations, along with overall results from device- and circuit-level perspectives. In Chapter 3, we perform a similar evaluation for JL-FETs targeted to meet the same nodes using similar FinFET designs. We then compare the results to those presented in Chapter 2 to assess whether JL technology poses a fundamental advantage or disadvantage over IM technology in terms of performance and manufacturability for near-term

generations. In Chapter 4, we extend the assessment to TFETs and identify whether tunnel-based devices behave any differently to IM- or JL-FETs, and if their vulnerability to LER or RDF begs certain design implications for future adoption. In Chapter 5, we investigate whether or not the contributions of LER and RDF can be treated independently or if they show any interaction in IM-FETs, JL-FETs, and TFETs. In Chapter 6, we examine how band structure and degeneracy effects in III-V materials such as InGaAs lead to differences in electronic response to RDF variability for JL-FETs, and if this implies that low density of states materials possess an inherent advantage over silicon from a variability standpoint. We experimentally demonstrate the cointegration of III-V JL-FETs on silicon substrates via nanotransfer printing in Chapter 7 and develop a framework to evaluate the potential benefits of heterogeneous integration technology in the context of realizable performance gains under manufacturability constraints. Finally, in Chapter 8 we demonstrate a process technique to fabricate on-chip supercapacitors onto silicon substrates with the potential for heterogeneous integration with CMOS circuits. Furthermore, a simulation framework is proposed to enable design exploration and optimization of on-chip supercapacitors which, along with our fabrication process, can serve as a launching point for bringing microscale supercapacitor technology to the domain of general micro- and nanoelectronics.

The long-term impact of this study will help steer the semiconductor industry in the direction of emerging device technologies which are best suited for near-future commercial adoption and, perhaps more importantly, spearhead heterogeneous integration as an alternative philosophy for improving circuit and system performance beyond the concepts of physical scaling and related challenges from variability.

Chapter 2

Inversion-Mode Silicon FinFET Variability

2.1 Background

Multi-gate FinFETs [6] have already entered commercial production starting at the 22nm node (year 2011) [5] and will likely become the standard FET architecture for years to come. The FinFET, as shown in Fig. 1, is a natural extension of the classic metal-oxide-semiconductor field-effect transistor (MOSFET) structure which arose from years of scaling efforts to maintain adequate electrostatic gate control over the channel. Here, the semiconductor channel is "folded" into a vertical stripe which can be defined via standard lithography with a resist mask, or by a sacrificial spacer mask. By wrapping the gate electrode over both sides of the (normally intrinsic) fin-shaped channel, the device maintains excellent short channel effect (SCE) control and high current drivability along the fin sidewalls [7]. Normally, a tri-gate (TG) FinFET with three conducting surfaces

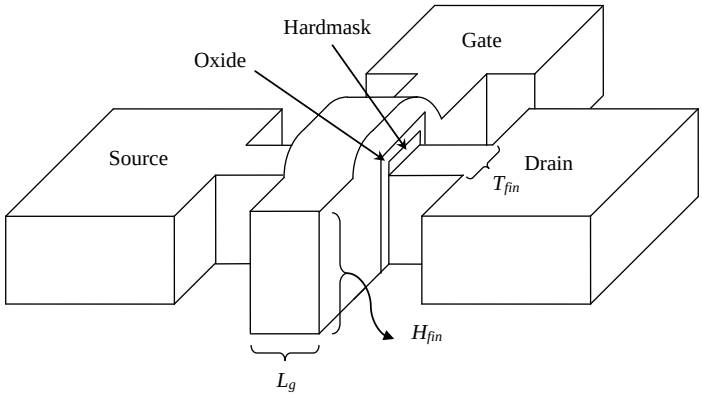


Fig. 1. Schematic of the FinFET structure. A vertical fin made up of silicon is straddled by a metal gate layer running perpendicular to the fin length. The metallurgical gate length L_g and the fin body thickness T_{fin} are indicated in the diagram. The effective channel width is the sum of T_{fin} and $2 \times H_{fin}$, where H_{fin} is the fin height. If the gate oxide covering the top surface of the fin is much larger than on the sidewalls, then the structure resembles a simple double-gate MOSFET and the channel width can be approximated as simply twice the fin height.

will result when the gate electrode is directly deposited and patterned after gate oxidation. If the insulator thickness between the gate and the channel is much thicker on the top surface (due to the presence of a hardmask) than the fin sidewalls, the top fin surface is negligibly driven by the gate and a double-gate (DG) FinFET results. TG FinFETs are generally easier to fabricate than DG FinFETs because a relaxed fin thickness T_{fin} to fin height H_{fin} ratio (typically $T_{fin}/H_{fin} \cong 1$) can be used to achieve the same SCE control as a DG-FinFET ($T_{fin}/H_{fin} \geq 5$). However, corner effects also become important in TG FinFETs whereas they are negligible in DG FinFETs [8].

Owing to the similar operation between FinFETs and planar MOSFETs, as well as their shared compatibility with standard CMOS processing, guidelines for FinFET scaling are mostly similar to those for planar devices. The major differences, however, are related to the values of T_{fin}, H_{fin}, and the fin pitch P_{fin}. In SOI devices, the gate length to body thickness ratio L_g/T_{body} must be kept sufficiently high (about 3:1) to maintain good SCE control, while the channel width W is continuously variable. In MG FinFETs, however, the ratio T_{fin}/L_g can be relaxed to roughly 3:2, since the fin body is equally divided between two gates. Ideally, T_{fin} should be as small as possible (especially for DG FinFETs) within processing constraints to allow consistent scaling of L_g for smaller nodes. In FinFETs, H_{fin} (analogous to the channel width W in planar devices) becomes fixed, forcing circuit designers to place multiple fins in parallel to increase the current drive by discrete multiples of H_{fin}. A lower value of H_{fin} reduces the quantization effect but demands more layout area for a given design, making the choice for H_{fin} a tradeoff. Since multiple fin designs are usually required in circuit applications, P_{fin} should be kept small enough to ensure that FinFET designs remain competitive against planar CMOS in terms of performance versus area. A simple analysis shows that $P_{fin} \leq 2H_{fin}$ represents the upper limit for the fin pitch assuming equal current

drive between FinFET and planar CMOS technologies, and neglecting additional area overheads [9].

2.2 IM-FinFET Modeling

The all-silicon IM-FinFET devices modeled in this study are generated using commercial TCAD software by Synopsys Sentaurus [10] and are shown in Fig. 2. They represent true SOI DG FinFETs where it is assumed that the hardmask in Fig. 1 and the buried oxide are both infinitely thick, which means the structure can be modeled via 2-D simulations. With this simplification, current transport is entirely parallel to the wafer plane. Specific design parameters for the IM-FinFETs are given in the upper portion of Table 1 which are targeted to meet the 2009 ITRS [11] guidelines for high-performance logic at 32, 21, and 15nm nodes. The baseline performances of each IM-FinFET generation are given in the lower portion of Table 1 for six different metrics:

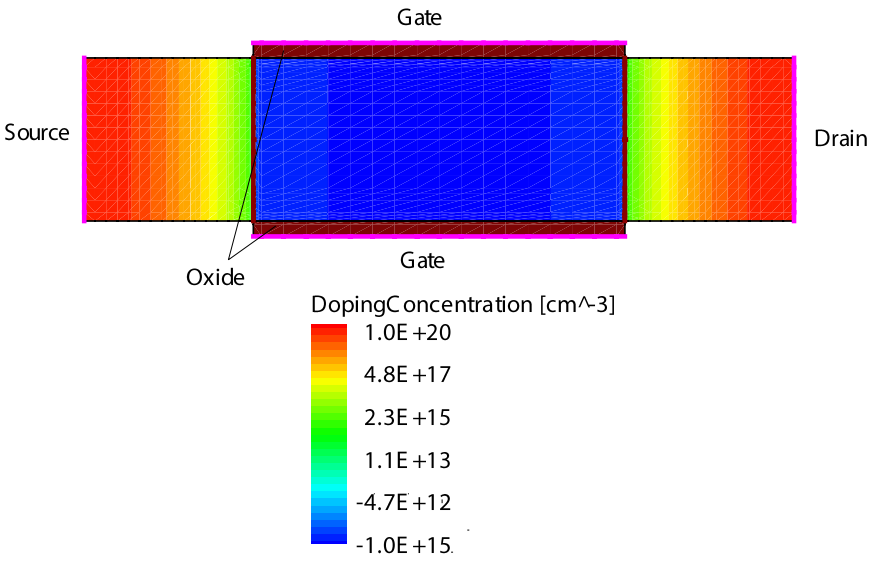


Fig. 2. Schematic of the 2D structure used to model IM-FinFET devices (32nm case shown). The structure represents a planar cut across the fin height and parallel to the plane of the wafer.

Table 1. Nominal Parameters for Simulated IM-FinFETs

Quantity	Technology Node			Description
	32nm	21nm	15nm	
L_g (nm)	22	17	13	Physical gate length
EOT (nm)	0.90	0.77	0.64	Equivalent oxide thickness
N (cm^{-3})	10^{15}	10^{15}	10^{15}	Body/fin doping
T_{fin} (nm)	9.6	8	6.4	Fin thickness
L_{sp} (nm)	10	8	6	Spacer width
Ψ_M (eV)	4.47	4.47	4.47	Gate work function
V_{DD} (V)	0.9	0.81	0.73	Power supply voltage
$V_{T,lin}$ (mV)	272	282	298	Lin. threshold voltage (max g_m method with V_{DS} = 50 mV)
$V_{T,sat}$ (mV)	201	203	208	Sat. threshold voltage (constant $I = W/L_g \times 10^{-7}$ A with $V_{DS} = V_{DD}$)
I_{on} (μA/μm)	1432	1527	1734	On-state drive current with $V_{GS} = V_{DS} = V_{DD}$
I_{off} (nA/μm)	6.7	9.7	13.3	Off-state leakage current with $V_{GS} = 0$ & $V_{DS} = V_{DD}$
SS (mV/dec)	67.9	69.8	71.6	Subthreshold swing
DIBL (mV/V)	24.0	32.0	39.7	Drain-induced barrier lowering

linear and saturation threshold voltage ($V_{T,lin}$ and $V_{T,sat}$), on-state drive current (I_{on}), off-state leakage current (I_{off}), subthreshold swing (SS), and drain-induced barrier lowering (DIBL). These numbers serve as the reference by which we normalize any LER/RDF variations to. Only *n*-type devices are considered in this work, where it is assumed *p*-type devices will only differ in their nominal performance levels with no impact on variability trends.

Sentaurus Device is used to simulate device behavior with a number of physics models activated to account for various phenomena which will be described next. A calibrated hydrodynamic (HD) transport model is used to capture high-field transport in the near-ballistic regime, giving a reasonable balance between accuracy and simulation time. In our work, the electron energy relaxation time τ_n = 1.4 ps and flux coefficient r_n = 0.3 are used based on calibrations done against Monte Carlo (MC) simulations performed by [12]. Quantization effects, which shift the peak carrier concentrations away from the oxide-channel interface and result in volume inversion, are taken into account using the density gradient approximation (DGA) with default parameters for Si. Again, this gives a good balance between accuracy and speed which is crucial for statistical variability studies. Mobility degradation due to impurity scattering, surface roughness scattering,

and HD transport are also considered using the Masetti [12], Lombardi [15], and Canali models [16], respectively.

2.3 Line Edge Roughness Modeling

Line edge roughness is a stochastic (i.e., random) variability mechanism, and hence its characterization requires a statistical description. Typically, rough line edge patterns are described by two parameters: the root-mean-square (rms) roughness amplitude σ_{LER} and the correlation length λ. Often times the $3\sigma_{LER}$ value is implied when one refers to "LER" in the literature; in this work, however, "LER" will refer to the one-sigma rms value. If the roughness amplitude on both edges of a line pattern are equal, then the LWR is related to LER by

$$\sigma_{LWR}^2 = 2\sigma_{LER}^2(1 - \rho_X) \tag{1}$$

where ρ_X is the cross-correlation coefficient between the two edge patterns. Standard resist patterning typically generates uncorrelated edges, i.e., $\rho_X = 0$, whereas spacer patterning produces correlated edges, i.e., $\rho_X = 1$ (ideally) which yields $\sigma_{LWR} = 0$. Modeling these two limits allows us to cover the entire range of LER-LWR cross-correlations made possible in a given process technology, and permits us to obtain variability estimates for fractional values of ρ_X by simple interpolation from the limiting data.

LER patterns are generated using the 1D Fourier synthesis approach described in [17] which involves taking the inverse Fourier transform of a known power spectrum $S(k)$ corresponding to some autocorrelation function (ACF), usually either in a Gaussian or exponential form. Random phases are applied to each component of the power spectrum to ensure that each pattern is unique and random. Thus for each desired combination of σ_{LER} and λ, a sufficient number of uniquely random LER patterns can be generated and used as inputs to the simulated FinFETs.

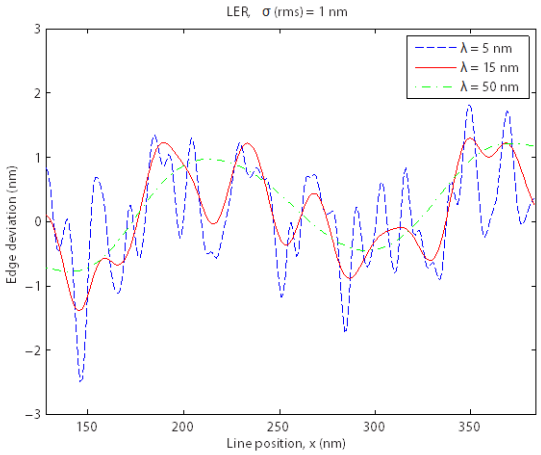


Fig. 3. Gaussian LER patterns corresponding to σ_{LER} = 1 nm and λ = 5, 15, and 50 nm. Short segments of these patterns were used as inputs to the simulated FinFETs.

Examples of Gaussian LER patterns generated with a MATLAB® script having σ_{LER} = 1 nm and λ = 5, 15, and 50 nm are shown in Fig. 3. Longer correlation lengths result in LER peaks and valleys being separated by larger distances, even though the roughness amplitude remains constant. The Gaussian LER model is chosen for reasons which will be explained next. Surface smoothing treatments such as thermal annealing [18], [19], sacrificial oxidation [20], and resist trimming [21] are capable of eliminating the majority of high-frequency roughness, leaving mostly low-frequency roughness in etched features. Moreover, it has been shown [22] that low-frequency roughness is the more significant source of intra-die variability characteristic of LER. With this in mind, we desired a simple analytical form for the ACF in order to reduce the simulation complexity, and one whose power spectrum consisted of mostly low-frequency roughness and negligible contribution from higher frequencies, leading us to consider the Gaussian model over the exponential model which retains non-negligible high-frequency components.

In our simulations, we consider the LER range $0 \leq \sigma_{LER} \leq 1$ nm to represent typical LER values which may be required by industry heading beyond 32nm technology, based on the 2009 ITRS [11] forecast and experimental data [17]. We fixed the correlation length at λ = 15 nm in our

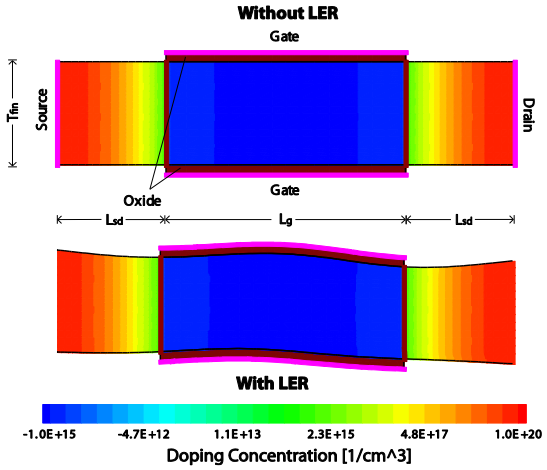


Fig. 4. Simulated IM-FinFET structures with and without 1 nm LER along the fin sidewalls.

work for several reasons: 1) to limit the permutations of σ_{LER}, λ, and technology node to a reasonable number in our study; 2) previous studies [22], [23] have shown that the effect of λ diminishes as λ > 15–20 nm; and 3) some experimental data has shown that current values of λ are estimated between 20–30 nm [17] and generally reduces with technology, suggesting λ = 15 nm as a reasonable estimate for sub-32nm generation lithography. An example IM-FinFET with and without LER applied to the fin edges is depicted in Fig. 4; this represents the case of "fin LER". We can see that fin LER results in fluctuation of the fin/body thickness along the channel direction, which can alter the transistor's SCE control and subsequent performance. While LER could also be present on the gate edge in an actual FinFET, representing "gate LER", previous studies have shown the effect to be less detrimental than fin LER at the 32nm node [22]. For this reason, only fin LER will be considered in the remainder of this study with the understanding that gate LER will not be as critical, especially for smaller nodes where the relative importance of fin LER over gate LER becomes even more apparent. Henceforth, we will also take any mention of "LER" to mean fin LER, unless indicated otherwise.

2.4 LER-Induced Variability in IM-FinFETs

The LER impact on lithographically-defined ("resist") IM-FinFETs is shown in Fig. 5 as a function of σ_{LER}. Moderate to large variation of $V_{T,lin}$ and $V_{T,sat}$ with LER is evident, especially in the latter case where $\sigma V_{T,sat}$ can exceed 10%. As expected, the 15nm devices show the most variation, while the 32nm devices show the least. The threshold voltage variation depends linearly on σ_{LER} since the total depletion charge in a fully depleted FinFET is directly impacted by fin thickness fluctuations, i.e., fin LER. This amount of LER-induced V_T variation may be troublesome in circuits requiring precise threshold voltage matching. Similar levels of V_T variation due to fin LER have also been found in [22] and [24].

I_{on} variation exhibits a similar but weaker dependence considering that σI_{on} can easily be kept within 10% of the nominal value in each technology node up to σ_{LER} = 1 nm; similar findings have also been reached in [22] and [24]. Note that σI_{on} is also linear with σ_{LER} since drive current is linearly proportional to $V_{DD} - V_T$ in velocity saturated FETs. The variation of I_{off} is much more

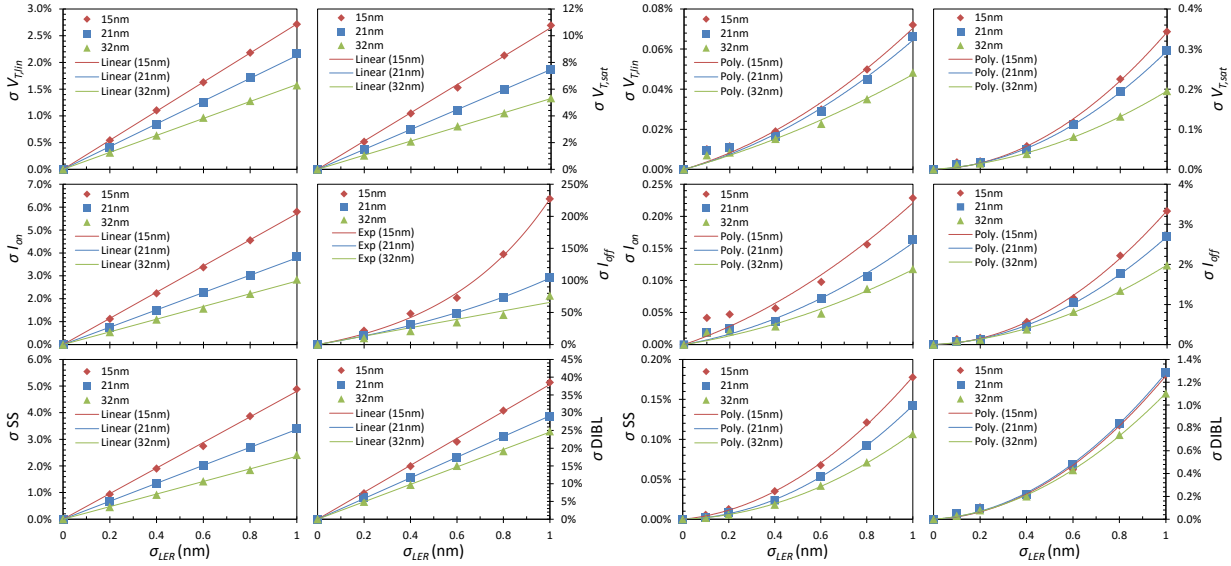


Fig. 5. Resist (left) and spacer (right) IM-FinFET device variability as a function of LER amplitude and technology node. Markers indicate actual simulated data while solid lines indicate best fits. Note the zoomed scale for spacer IM-FinFET data compared to resist IM-FinFET data.

pronounced, however, where σI_{off} varies exponentially with σ_{LER} (since I_{off} is an exponential function of V_T) and reaches more than 200% of the nominal value for 15nm devices. Such wild fluctuations in I_{off} may be detrimental to circuit performance if the power dissipation of individual devices and circuit blocks cannot be kept within acceptable margins. In light of these results, it appears that the drastic variation of I_{off} due to fin LER may be a critical obstacle toward further scaling of FinFETs beyond 32nm.

The effect of LER on SS is somewhat low on the order of a few percent, and is also linear since the fluctuation of T_{fin} due to $\sigma_{LER} \leq 1$ nm can be treated as a linear perturbation in C_D, i.e., $C_D = \epsilon_{Si}/(T_{fin} + \Delta T_{fin}) \approx \epsilon_{Si}/T_{fin}(1 - \Delta T_{fin}/T_{fin})$, where ΔT_{fin} is roughly given by σ_{LER}. DIBL variation is more considerable—σDIBL easily exceeds 10% in each generation over the LER range—as opposed to the SS variation, which can be kept under 10% for the entire LER range.

For spacer-defined ("spacer") IM-FinFETs, the impact of LER is drastically reduced in terms of parameter fluctuations for all three technology generations. Note the zoomed vertical scales used in Fig. 5 for spacer IM-FinFETs compared to those for resist IM-FinFETs. From the data, the elimination of LWR by spacer lithography (due to sidewall correlation) offers substantial improvement in minimizing device variation. These results compare well to the findings in [22] which demonstrate a significant reduction in the saturation threshold voltage mismatch and current factor mismatch to less than 1% of the nominal values over a similar LER range. We also observe that in most cases the variability curves show less dependence on the actual technology node for spacer IM-FinFETs. In other words, there is little difference between the 32, 21, and 15nm cases here. From this, we see that the presence (absence) of LWR is responsible for the observed variability trends in the resist (spacer) IM-FinFETs, rather than the actual LER itself.

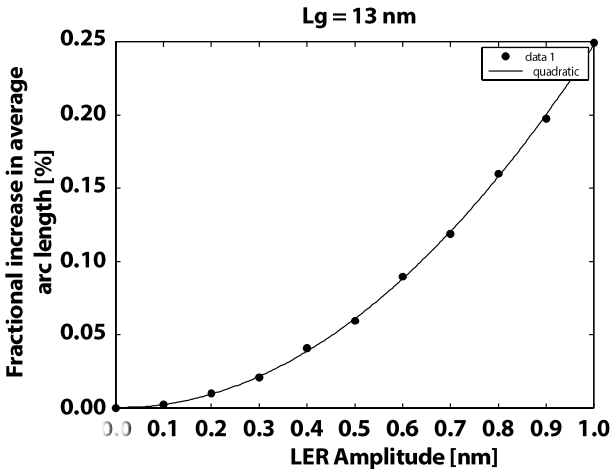


Fig. 6. Quadratic rise in average arc length for spacer FinFETs due to LER as a function of root-mean-square amplitude. The nominal arc length corresponds to a 13 nm channel length for the data shown.

Interestingly, every parameter investigated appears to vary quadratically, rather than linearly, with σ_{LER}. To explain why, we first observe that because of the correlated fin edges in a spacer IM-FinFET the body thickness does not change along the length of the fin, i.e., $\sigma_{LWR} = 0$. However, the presence of LER causes the body/channel region to bend and curve in shape which results in a curved potential profile compared to an ideal device, and hence, the path for current should roughly follow the curvature of the fin geometry. Mathematically, the total arc length from source to drain can only lengthen due to random vertical displacement of the fin edge, i.e., LER, and the fractional increase in arc length tends to increase quadratically with the root-mean square vertical deviation. This was confirmed by directly analyzing the LER patterns in MATLAB and determining the relationship between average arc length and roughness amplitude as shown in Fig. 6. Variation in the arc length due to LER can thus be treated as variation in the effective channel length of the device which is subsequently manifested in the trends of Fig. 5.

Note that we have assumed perfectly correlated fin sidewalls, i.e., zero LWR, in this analysis. In reality, spacer lithography may not generate 100% correlated edges on both sides due to variations in the deposition and etch processes, or subsequent annealing steps. Experimentally, it

has been shown that the actual LWR can be nonzero in spacer-defined FinFETs [22] so that a more realistic estimate of spacer FinFET variability would likely involve a weighted average of the resist and spacer FinFET results, where the emphasis on each depends on the magnitude of the cross-correlation coefficient ρ_X. However, systematically generating random LER patterns where each top-bottom pair represents a deterministic ρ_X is nontrivial and impractical here.

Ultimately, the impact of LER does not appear to pose a major obstacle for IM-FinFETs to meet the demands of future generations (15nm and smaller) given current and projected lithography capabilities (i.e., $\sigma_{LER} \leq 1$ nm). This finding is primarily attributed to the robustness of IM-based technology whose fundamental mode of operation is not jeopardized by geometric fluctuations arising from LER. In IM-FETs, switching is predicated on the existence of opposing *p-n* junctions at the source-channel and drain-channel interfaces to block current flow in the "off" state, and bridging those junctions by means of electrostatically generating an inversion layer which enables current flow in the "on" state. Fundamentally, this action has no outright dependence on the geometry (i.e., thickness) of the channel, meaning any geometric fluctuations within the channel do not directly prevent the switching operation from happening. Only at short channel lengths does the body thickness matter, yet it remains a secondary effect only. It is for this reason that IM-FinFETs remain viable at small geometries even in the presence of LER, whereas other FET technologies (e.g., junctionless) may not, which we will discover later.

2.5 Random Dopant Fluctuation Modeling

Random dopant fluctuation is another stochastic variability mechanism, and hence, is described statistically. Unlike LER, however, truly quantitative measurements of RDF are extremely

difficult to achieve due to the high difficulty in mapping the exact position and number of individual dopant atoms in actual devices. This differs from LER, in which case the position of line edges can be easily visualized and characterized with critical dimension scanning electron microscopy (CD-SEM). As a result, details on the microscopics of RDF are not easily measured experimentally, even if the macroscopics are well understood. For example, extracting the macroscopic doping concentration in a large semiconductor sample is relatively straightforward, but locating the individual dopant atoms with atomic precision is not. Based on statistical theory, however, we can easily calculate the probability that a given lattice site will contain a substitutional impurity, assuming a completely random occupational process, since the product of the occupation probability and the semiconductor volume (which is easily obtained) must be equal to the macroscopic doping concentration (which can be electrically measured). The resulting probability follows a Poisson distribution, and sufficiently describes the purely random nature of RDF, neglecting any systematic variation sources (e.g. specific variations in implanted dose, angle, energy, etc.). While imperfect, this treatment gives us a simple and effective way to treat RDF in real devices.

In our simulations, we adopt the aforementioned approach in Sentaurus TCAD by randomizing the position and number of dopant atoms in a doped semiconductor region according to a Poisson distribution such that, on average, the total number of dopants added to a device is equal to the integrated macroscopic doping profile(s) over the entire device volume. This gives a random total number of dopants N_T to each device such that $\sigma N_T = N_T^{1/2}$ where $N_T = N \times (volume)$. In other words, regions with higher macroscopic doping concentration N have a larger fluctuation in total number of dopants. When the fluctuation is normalized to the nominal N_T, however, we obtain $\sigma N_T/N_T = N_T^{-1/2}$ so that highly scaled devices with low N_T experience stronger RDF effects. This

properly reflects the situation where RDF becomes more problematic for finer technology generations.

Once the proper number of dopants are placed randomly within a device, the resulting electrostatic impact is determined by the procedure proposed by Sano *et al* [25], in which the Coulomb potential associated with individual dopants is decomposed into long-range and short-range components, separated by a screening length $1/k_c = 2\times N(x,y,z)^{-1/3}$ where $N(x,y,z)$ is the impurity concentration in a unit mesh volume centered at (x,y,z). Only the long-range component is retained by the device simulator in drift-diffusion simulations in order to avoid pitfalls associated with the short-range component which generates unrealistic potential singularities at mesh nodes containing an impurity. Note that the short-range potential component is not unrealistic in and of itself; the problem is that it is incompatible with traditional DD-based simulators when stretched to the atomistic limit—it predicts that mobile carriers become trapped near the ionized centers, resulting in excessive screening of the dopants and an underestimation of the depletion charge. Once the potential profile is calculated with RDF according to the above procedure, an effective doping concentration is created which varies by location and models the effect of RDF during

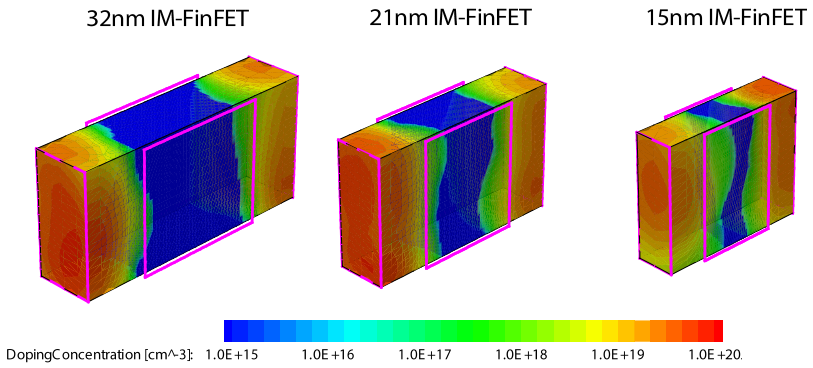


Fig. 7. Effective doping profiles resulting from RDF in 32, 21, and 15nm IM-FinFET devices. The effective channel length becomes nonuniform and reduces on average at smaller nodes.

actual device simulations. Fig. 7 shows the effective doping concentration in three IM-FinFET generations when RDF is considered in the source and drain regions. Here, the effective channel length L_{eff} becomes nonuniform because of the random location and number of dopants near the source/drain-gate edges. Because of the dependence on volume, 3-D simulations are required for RDF analysis instead of just 2-D for LER.

2.6 RDF-Induced Variability in IM-FinFETs

By virtue of its intrinsic channel, IM-FinFETs are widely believed to be immune to RDF, only except when an occasional dopant appears as a "contaminant" [26]. From a quick inspection of Fig. 7, we see this is not completely true since, at very small metallurgical gate lengths, the effective channel length experiences sizeable fluctuation as mentioned previously. Thus, while the channel remains completely intrinsic, RDF in the source and drain may become important when scaled to nanometer dimensions.

A simple scaling law which describes performance mismatch due to stochastic variability in planar MOSFETs was formulated by Pelgrom [27], which states that the variation in a given parameter, such as V_T, can be expressed as

$$\sigma V_T = \frac{A_{V_T}}{\sqrt{W \times L}} \tag{2}$$

where A_{V_T} is some coefficient, and $W \times L$ represents the active area of the device. In other words, larger sized devices exhibit less performance variation due to self-averaging of (stochastic) variability effects, with an inverse relationship appearing. The mismatch coefficient describes the relative sensitivity to process variation for a device, and has units of mV·μm in the above case. While

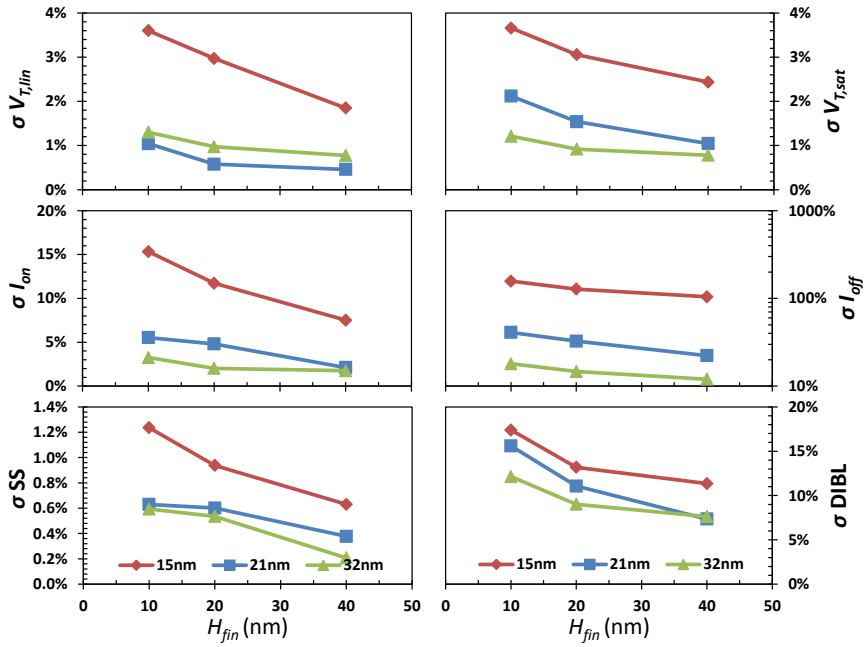


Fig. 8. RDF-induced variability in IM-FinFETs as a function of fin height and technology node.

originally intended to describe variability in planar MOSFETs, we may anticipate the same trend to apply to FinFETs with H_{fin} replacing W in the previous equation.

The RDF impact on IM-FinFETs is presented in Fig. 8 which shows that Pelgrom's scaling law also applies to FinFETs as well. As H_{fin} increases, performance variation generally reduces with an inverse relationship appearing, and smaller technologies such as 15nm tend to exhibit more RDF-induced variability than larger technologies such as 32nm. We also observe that $\sigma V_{T,lin}$ and $\sigma V_{T,sat}$ are kept below 5% in all cases, σSS is kept below 2%, and σDIBL below 20%, all of which are good results and demonstrate the advantage of having a completely intrinsic channel to suppress RDF. Leakage current variation is also relatively well controlled with $\sigma I_{off} \leq 100\%$, compared to roughly 200% with 1 nm LER. However, σI_{on} is still somewhat large reaching up to 15% for 15nm IM-FinFETs; this is a direct result of the shortened, and highly variable, L_{eff} resulting from RDF illustrated in Fig. 7. In this respect, the major concern for RDF in IM-FinFETs will likely be variation in drive current, especially as the gate length is scaled toward nanometer dimensions. We

should note, however, that RDF does not jeopardize the intrinsic switching capability of IM-FinFETs since the electrostatic generation and removal of an inversion layer in the channel occurs regardless of whether RDF exists in the source and drain. RDF only changes L_{eff} and as such, it has a secondary impact on transistor performance only—again, this will be in contrast to the situation for JL-FETs. Overall, however, IM-FinFET technology demonstrates good resistance to RDF by virtue of its intrinsic channel with minimal impact on threshold voltage and SCE control.

2.7 Circuit-Level Variability Impact[1]

Besides focusing on device-level results to evaluate the performance of IM-FinFET technology in the presence of variability, we should also examine the resulting circuit-level impact to better understand how the performance of representative circuits will be affected in real applications. To do this, we developed a framework to adapt the device-level variability data obtained from TCAD simulations to be implemented in circuit simulations of large-scale microprocessors

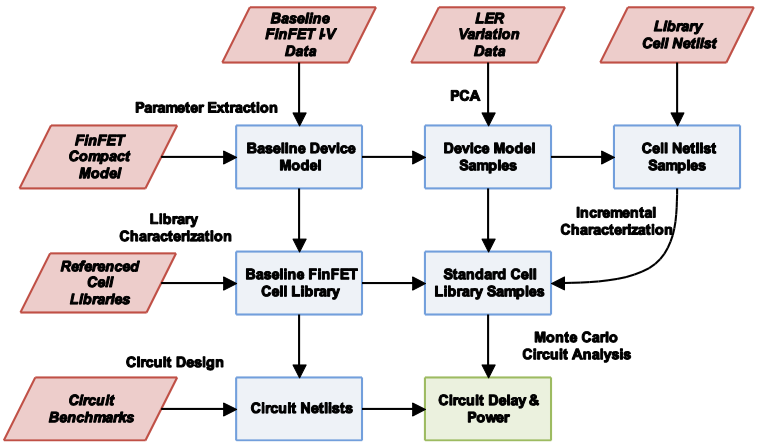


Fig. 9. Overall flow of the circuit benchmark evaluation process.

[1] We sincerely thank Liangzhen Lai and Prof. Puneet Gupta from the UCLA NanoCAD Laboratory for their valuable contributions to this section.

for delay and power analysis. The overall process flow is illustrated in Fig. 9 and can be briefly summarized in the following steps:

1. A reference FinFET compact model [28] is fitted to match the nominal *I-V* characteristics obtained from TCAD simulation using parameter extraction to generate a baseline compact model.
2. The baseline compact model is then used to characterize a baseline cell library that contains the timing and power information of each logic gate, which will later be used for circuit synthesis, placement and routing (SPR) and further incremental characterizations. SPR is performed for two processor benchmarks: MIPS [29] and ARM Cortex-M0 [30], clocked at different periods (fast, medium, and slow).
3. Variability is modeled by varying the compact model parameters such that device metric sample variations match with those obtained from device-level TCAD simulation. The method of principal component analysis [31] is used to translate device-level variations to compact model parameter variations [32].
4. Using the compact model samples, the cell library samples are then generated from our baseline library and incrementally characterized to simulate their resulting circuit performance by conventional tools.
5. Finally, Monte Carlo circuit analysis is performed on the cell library samples to extract variations in delay and power.

A summary of the findings for the LER impact on IM-FinFET circuit performance is given in Table 2 which shows the normalized mean increase and variation for circuit delay and leakage power in 32, 21, and 15nm resist and spacer IM-FinFETs with 1 nm LER. Despite the nonnegligible device-level variations from 1 nm LER in Fig. 5, negligible (< 1%) variation is observed for

Table 2. Delay and Leakage Mean and Sigma over All Benchmarks with 1 nm LER for Resist (R) and Spacer (S) FinFET Technologies

Node	Delay			Leakage		
	Baseline w/o LER	Mean w/ LER	Sigma w/ LER	Baseline w/o LER	Mean w/ LER	Sigma w/ LER
32-S	952 ps	100%	0.00%	14.65 μW	100%	0.0%
32-R		101%	0.15%		114%	0.1%
21-S	635 ps	100%	0.00%	11.47 μW	100%	0.0%
21-R		106%	0.30%		125%	0.1%
15-S	381 ps	100%	0.01%	6.59 μW	100%	0.0%
15-R		102%	0.04%		149%	0.2%

delay and leakage across all benchmarks in all IM-FinFET technologies considered. This conclusion is attributed to the stochastic nature of LER, whose effects average out between different cells along a critical path. We also observe that there is negligible increase in mean delay due to LER while up to 49% increase in mean leakage is obtained for 15nm resist IM-FinFETs. The increase in mean leakage power is caused by a mean increase in I_{off} due to LER from the device level, while the lack of mean delay change results from no discernible increase in I_{on} from LER at the device level. Overall, the impact of LER on large-scale digital microprocessors is minimal, except for a moderate increase in mean leakage power for resist IM-FinFET technologies. Additionally, spacer lithography eliminates all LER impacts at the circuit level. These conclusions will likely be repeated even for RDF considering the purely stochastic nature of both variability mechanisms.

2.8 Summary

The impact of LER on IM-FinFET variability is well-managed at the 32, 21, and 15nm nodes, with smaller nodes (e.g., 15nm) exhibiting more variation than larger nodes (e.g., 32nm). (Fin) LER in IM-FinFETs results in nonuniform fluctuation of the fin/body thickness in individual devices, which primarily affects transistor SCE control. Resist-defined IM-FinFETs exhibit linear performance variation (except for I_{off}) versus LER amplitude, and fluctuation in most performance metrics are kept to a reasonable level. A maximum of 10% $\sigma V_{T,sat}$ is obtained for 15nm resist IM-

FinFETs at σ_{LER} = 1 nm, and σI_{on} is kept below 10% for all technology generations, both of which may be considered acceptable for logic applications. Variation in leakage current increases exponentially with σ_{LER}, with up to 250% σI_{off} obtained for 15nm resist IM-FinFETs at σ_{LER} = 1 nm. This is accompanied by an increase in mean leakage current which must be considered in a circuit's power budget design. For all performance metrics, the adoption of spacer lithography significantly alleviates the variability impact by eliminating LWR in individual devices, and results in quadratic variability trends with σ_{LER}. From a standpoint of variability management, spacer lithography will be an indispensable manufacturing option for future generations.

The impact of RDF on IM-FinFET variability is also well-managed, with smaller nodes again exhibiting more variation compared to larger nodes. Variability scaling in accordance with Pelgrom's law is observed with inverse dependencies appearing as a function of H_{fin}. Less than 5% $\sigma V_{T,sat}$ is demonstrated while up to 15% σI_{on} is obtained for 15nm IM-FinFETs with H_{fin} = 10 nm, highlighting the different impacts from LER and RDF. RDF in the source and drain causes fluctuation in L_{eff} and becomes more significant for highly scaled generations with smaller nominal L_g.

Both LER- and RDF-induced variability impacts are considered secondary for IM-FinFETs and should not pose significant problems for near-term (15nm and below) technology adoption given current and projected manufacturing capabilities. On this basis, IM-FinFETs remain a viable option for continued scaling in the presence of manufacturing variability.

Chapter 3

Junctionless Silicon FET Variability

3.1 Background

Junctionless FETs [33]–[37] have quickly become a popular topic in recent years as a possible replacement for standard IM-FETs due to their simplified processing and comparable performance. Fundamentally, the only defining characteristic of JL-FETs is the absence of any *p*-*n* junctions between the source, channel, and drain regions, hence the name "junctionless". Since JL-FETs lack any junctions, the nominal doping concentration is typically designed to be uniform and homogeneous throughout the source, channel, and drain regions, making JL-FETs resemble gated resistors. A crucial benefit of this is the ability to bypass processing steps which normally plague IM-FETs related to ultra-shallow junction formation and downstream thermal budget management. JL-FETs may be conceived in any standard configuration based on planar or non-planar architectures, including SOI, DG or TG FinFETs, NW-based FETs, etc. An example of a NW-based JL-FET is illustrated in Fig. 10 alongside a NW-based IM-FET.

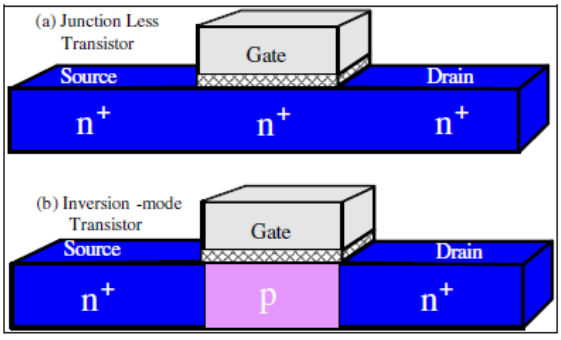


Fig. 10. Schematic comparison of (a) junctionless and (b) inversion-mode FETs and their associated doping profiles. From [36].

By applying different gate voltage values, the depletion region under the gate either pinches off the channel in the "off" state or opens up a buried channel in the "on" state. For an *n*-type JL-FET, the combination of a low gate voltage and an appropriate gate work function results in total depletion of the channel and an energy barrier for carriers between the source and drain. When a high gate voltage is applied, the depletion regions retract and the energy barrier vanishes, resembling flat-band (i.e., resistor-like) conditions in the channel. As a result, the JL-FET normally operates as a depletion-mode device rather than an IM device. The buried channel nature of JL-FETs is different from the surface channel nature of IM-FETs—this will have important consequences that will become apparent in later sections, especially when we discuss their performance vulnerability to LER and RDF.

3.2 JL-FET Modeling

The JL-FET structure modeled in this work resembles the same FinFET structure in Fig. 2, except for a different doping profile in the fin. In essence, the FET technology considered here will be JL-FinFETs designed to meet the same sub-32nm ITRS nodes for high-performance logic as before. This allows us to draw fair comparisons between the inherent advantages and disadvantages of JL and IM technologies when designed for the same physical layouts and operational targets. Thus, the theme for this chapter will be to answer the question: For the same physical specifications, operating conditions, and performance targets, which transistor technology best meets those deliverables with consideration of variability and manufacturing demands: inversion-mode or junctionless?

Table 3 lists the nominal parameters and performance metrics for the JL-FinFETs considered in this work. As in the last chapter, only *n*-type JL-FinFETs will be simulated in this study.

Table 3. Nominal Parameters for Simulated JL-FinFETs

Quantity	Technology Node			Description
	32nm	21nm	15nm	
L_g (nm)	22	17	13	Physical gate length
EOT (nm)	0.90	0.77	0.64	Equivalent oxide thickness
N (cm^{-3})	2×10^{19}	2×10^{19}	2×10^{19}	Body/fin doping
T_{fin} (nm)	9.6	8	6.4	Fin thickness
L_{sp} (nm)	10	8	6	Spacer width
Ψ_M (eV)	5.25	5.02	4.82	Gate work function
V_{DD} (V)	0.9	0.81	0.73	Power supply voltage
$V_{T,lin}$ (mV)	306	306	300	Lin. threshold voltage (max g_m method with V_{DS} = 50 mV)
$V_{T,sat}$ (mV)	200	192	185	Sat. threshold voltage (constant $I = W/L_g\times10^{-7}$ A with $V_{DS} = V_{DD}$)
I_{on} (μA/μm)	1144	1225	1330	On-state drive current with $V_{GS} = V_{DS} = V_{DD}$
I_{off} (nA/μm)	11.3	21.3	36.4	Off-state leakage current with $V_{GS} = 0$ & $V_{DS} = V_{DD}$
SS (mV/dec)	72.5	74.2	75.3	Subthreshold swing
DIBL (mV/V)	77.3	89.8	95.6	Drain-induced barrier lowering

Comparing these values with those of Table 1 for IM-FinFETs, the only design differences lie with the nominal doping concentration ($N = 2\times10^{19}$ cm^{-3} for JL-FinFETs) which is the same for the source, drain, and channel regions, and the gate work functions which are adjusted for each node to obtain $V_{T,sat} \cong 0.2$ V with $I_{off} < 100$ nA/μm according to the ITRS definition. The baseline performance values for JL-FinFETs remain comparable to those of IM-FinFETs, despite being marginally worse in all regards. I_{on} is about 20% worse and I_{off} is about 50% worse for JL-FinFETs compared to IM-FinFETs at the same generation, and JL-FinFETs consistently have higher SS and DIBL. These conclusions are well explained by the nature of buried channel formation in JL devices, resulting in weaker gate-channel capacitive coupling leading to worse SCE control. In a JL device, the highly doped channel results in significant mobility degradation due to impurity scattering despite the reduction in surface roughness scattering from the reduced transverse electric field above threshold. This is a necessary tradeoff to prevent excessive current loss from parasitic resistance if a lower channel doping were used. For IM-FinFETs, mobility degradation due to surface roughness scattering at maximum gate voltage is mitigated at the geometries considered due to volume inversion in the channel, resulting in overall higher channel mobility (Fig. 11) compared to JL-FinFETs; this may explain the higher I_{on} compared to JL-FinFETs. Nevertheless, the

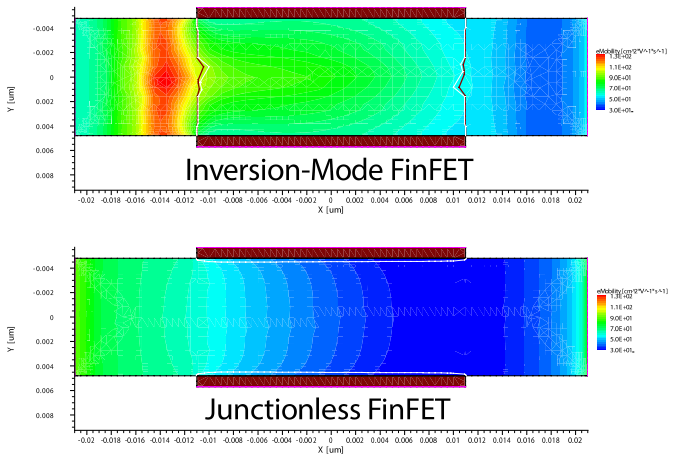


Fig. 11. Electron mobility plots in 32nm IM- and JL-FinFETs at V_{GS} = V_{DS} = V_{DD} = 0.9 V. The channel mobility is consistently higher in IM-FinFETs compared to JL-FinFETs due to reduced impurity and surface roughness scattering at these geometries.

comparable level of performance between JL- and IM-FinFETs should hold regardless of the actual FET architecture, especially if the operation of JL-FETs is extended beyond depletion to accumulation, thereby blurring the distinction between a strictly buried channel JL device and a hybrid buried-surface channel device. Recent experimental evidence from Intel [38] supports our conclusions—however we will still restrict the scope of our analysis to depletion-mode JL devices only.

The same set of device simulation models are used for our JL-FinFET simulations as they were for IM-FinFETs, namely the HD transport model, DGA for quantum corrections, and mobility models accounting for surface roughness scattering, doping dependence, and HD transport. Finally, device variability from LER and RDF are modeled in the same way for JL-FinFETs as they were for IM-FinFETs. Examples of 32nm JL-FinFETs with LER (σ_{LER} = 1 nm) and RDF (H_{fin} = 20 nm) are depicted in Fig. 12 and Fig. 13, respectively. Again, 2-D simulations are sufficient for LER analysis while 3-D simulations are required for RDF analysis.

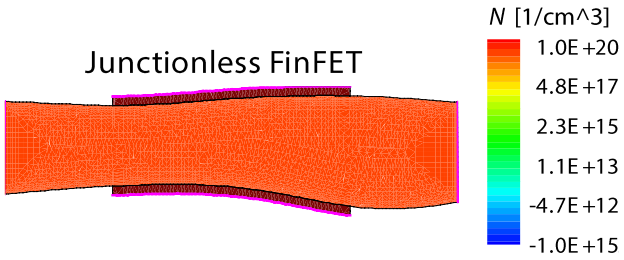


Fig. 12. Representative 32nm JL-FinFET with 1 nm LER applied to the fin edges.

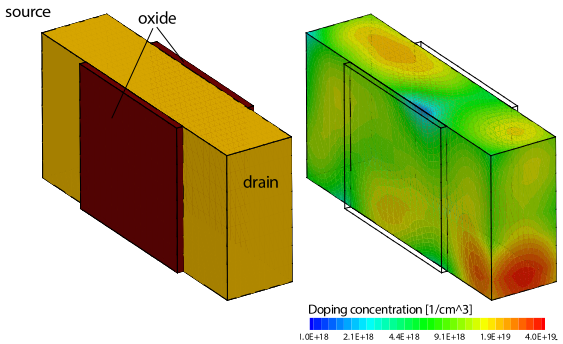


Fig. 13. Representative 32nm JL-FinFETs with and without RDF applied. The effective doping concentrations in both cases are shown with the same color legend.

3.3 LER-Induced Variability in JL-FETs

The extracted performance variations due to LER for resist-defined JL-FinFETs are shown in Fig. 14 as a function of σ_{LER} and technology node. The functional trends remain identical: linear versus σ_{LER} for all metrics except for σI_{off} which is exponential. Since JL-FETs, like their IM counterparts, are MOSFET-inherited designs, this is not surprising. We find, however, that JL-FinFET variability from LER (Fig. 14) is substantially worse than for IM-FinFETs (Fig. 5), especially in terms of $V_{T,lin}$, $V_{T,sat}$, I_{on}, and I_{off}. We see that $\sigma V_{T,sat}$ already exceeds 60% for σ_{LER} = 1 nm, by comparison this value was only 5–10% for IM-FinFETs. At the same LER amplitude, literature values for gate LER-induced $\sigma V_{T,sat}$ are in the range of 2–8% for L_g = 30 nm planar MOSFETs [17] and IM-FinFETs [22]. If one were to operate within a $\sigma V_{T,sat} \leq 20\%$ limit as suggested by the

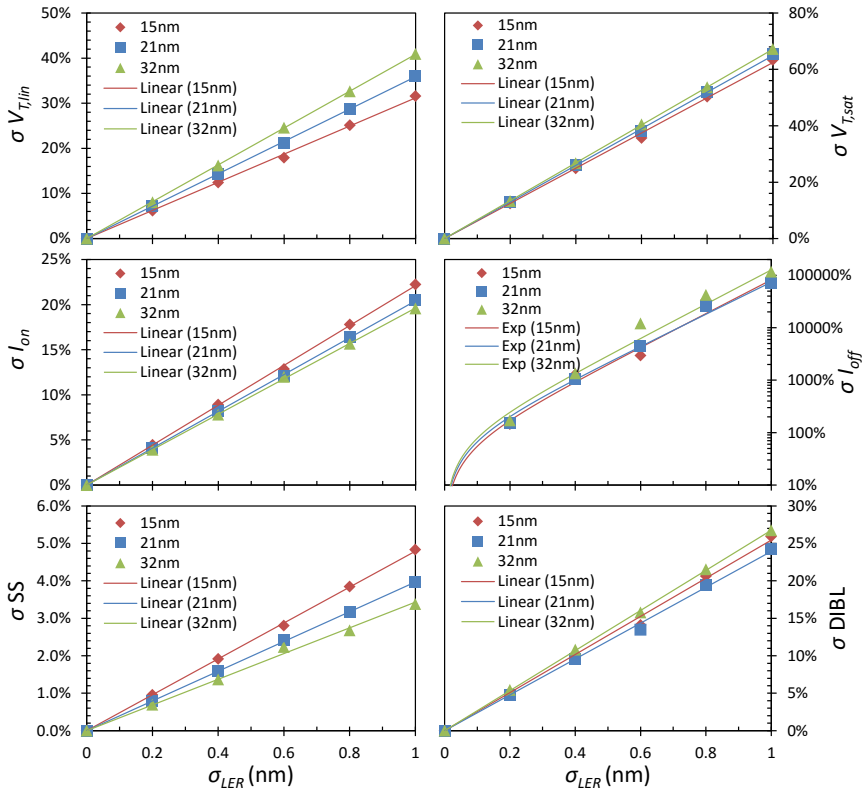


Fig. 14. Resist JL-FinFET device variability as a function of LER amplitude and technology node. Markers indicate actual simulated data while solid lines indicate best fits.

ITRS, then σ_{LER} would need to be kept at or below 0.2 nm—a major burden on state-of-the-art lithography. This partially agrees with data in [33] where JL-FETs fabricated on SOI wafers exhibiting surface roughness $\sigma T_{Si} \leq 0.2$ nm could expect $\sigma V_T = 20$ mV ($V_T \cong 0.2$ V). Additionally, with such high $\sigma V_{T,sat}$ values for JL-FinFETs it becomes very probable that some devices will have a negative threshold voltage. If only positive voltages are available, in some cases the extracted I_{off} no longer represents an "off-state" current. With this in mind, the actual leakage current values in Table 3 may be better interpreted as a minimum attainable current rather than a subthreshold current. Regardless, the key point is that LER has a significant impact on JL-FinFET variability, especially when compared against equivalent IM-FinFETs.

To explain these outcomes, we invoke a simple physical argument based on the operating principles of both device architectures. For traditional IM-FETs, a potential barrier exists at the

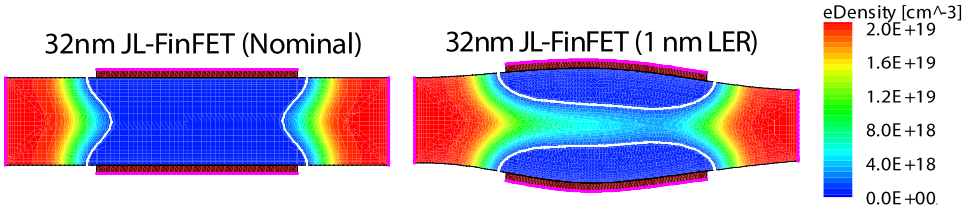


Fig. 15. Electron density plots for two representative 32nm JL-FinFETs showing the inadvertent formation of a conducting channel due to fin LER at V_{GS} = 0.1 V, and V_{DS} = 0. White lines indicate depletion region boundaries.

source-channel junction by virtue of the doping to prevent any significant current flow in sub-threshold. Ignoring short SCE for the moment, we observe that this is true no matter how thin or thick the body is, and hence, the existence of a potential barrier to subthreshold conduction is not threatened by body thickness variations due to fin LER. In fact, we surmise that long channel IM-FinFETs would exhibit negligible variability due to fin LER as suggested by the flatter 32nm curves compared to the 15nm devices in Fig. 14. However, IM-FinFETs still demonstrate some sensitivity to fin LER as the barrier magnitude is reduced via DIBL. Nevertheless, the existence of said barrier is still ensured by the presence of a junction, thus making it a robust device.

For depletion-mode JL-FETs, however, a barrier to subthreshold current is stipulated on the body/channel being fully depleted. Fig. 15 reveals how variations in the body thickness from fin LER may inadvertently cause a conducting channel to form near the midsection to overcome the barrier, thereby driving the transistor out of subthreshold. For long-channel devices, this conduit will be entirely responsible for the resulting LER-induced variability (which will be significant). For short-channel devices, the LER contribution from SCE will also add to the net variability. However, the similarity between the 32, 21, and 15nm JL-FinFET curves implies that the dominant mechanism is not acting through SCE as it was for IM-FinFETs, but instead through the inadvertent opening of a conducting channel. Furthermore, the fact that peak variations in SS and DIBL due to LER remain fairly similar between JL- and IM-FinFETs further indicates that SCE

degradation is not the major issue in JL devices. Rather, the (inadvertent) direct opening/closing of a conducting channel due to LER is reminiscent of a primary weakness in depletion-mode FETs: inherent dependence on a delicate balance between electrostatic and geometrical control. For this reason, LER has a primary effect on JL-FET variability because it directly jeopardizes transistor operation by negating proper geometrical control. This is in direct contrast with inversion-mode FETs which do not depend on geometrical control to ensure correct operation, and for which LER remains a secondary effect only. With this in mind, we see that by virtue of its characteristic nature, JL devices are not robust against LER whereas IM devices are, and its implications can be directly observed from Fig. 5 and Fig. 14.

Despite focusing on double-gate JL-FinFETs, the same behavior should occur for TG JL-FETs and JL-NWFETs operating as depletion-mode devices, based on similar conclusions [36] drawn from experimental data. Accumulation-mode JL-FETs may exhibit less sensitivity to LER with similar performance, but this remains to be seen. We also speculate that by adopting a TG design with a larger top gate, the sidewall fin LER impact may be alleviated at the expense of higher gate LER impact and layout area, as the device resembles more of a hybrid between vertical double-gate and traditional planar technology. Further investigation will be needed to judge whether this represents a favorable design tradeoff.

Finally, the reader may note that we have not considered the potential benefits of spacer lithography for reducing LER variability in JL-FinFETs. By correlating the fin sidewalls with spacer lithography, it stands to reason that elimination of LWR from JL-FinFETs would prevent accidental channel opening/closing (Fig. 15) from occurring since T_{fin} would no longer fluctuate along the device. Clearly, this would improve the situation for JL technology as a whole and re-

move many of the variability concerns from LER. At this point, we have not generated an exhaustive set of data to confirm this possibility; there remains room for further investigations to determine if spacer lithography can bring down JL-FinFET variability levels closer to those in Fig. 5 for spacer IM-FinFETs. As we will see in the next section, however, spacer lithography does not address the problem of RDF variability in JL-based devices, so spacer lithography only presents a partial solution at best.

3.4 RDF-Induced Variability in JL-FETs

Because of the high doping (N = 2×10^{19} cm^{-3}) and small device volumes used in our JL-FinFETs, the impact of RDF may become significant as the nominal dopant count ranges from ~300 (32nm with H_{fin} = 40 nm) to only 3 (15nm with H_{fin} = 10 nm). Fig. 16 shows the extracted performance variations of our JL-FinFETs with RDF versus H_{fin} and technology node. Several

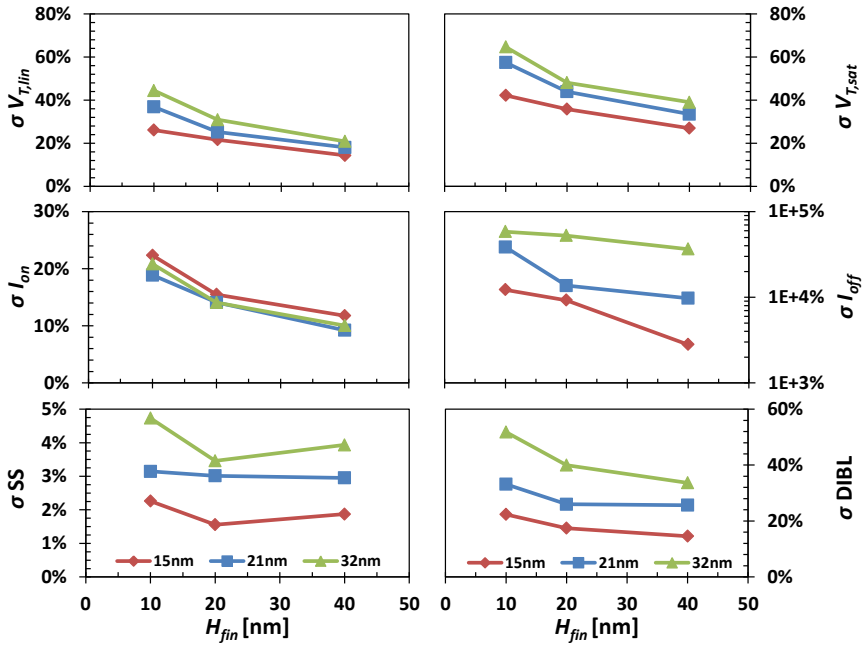


Fig. 16. RDF-induced variability in JL-FinFETs as a function of fin height and technology node.

observations are apparent: 1) the magnitude of device variability is quite large for all cases, especially in terms of $\sigma V_{T,lin}$, $\sigma V_{T,sat}$, σI_{on} and σI_{off}, 2) the curves show an inverse dependence with H_{fin} which correlates with Pelgrom's law, and 3) more aggressively scaled technologies tend to exhibit less variation with the exception of σI_{on}. The first observation is alarming, but not unexpected since V_T and parameters which depend on it are highly sensitive to the actual channel doping profile, and hence RDF. With $\sigma V_{T,sat}$ ranging between 20–60%, σI_{on} between 10–20%, and σI_{off} between 10^3–10^6 %, the effect of RDF is comparable to fin LER up to an amplitude of 1 nm. The second observation is consistent with traditional scaling of the "channel width", i.e. H_{fin} in a FinFET. The third observation is less obvious and cannot be explained by Pelgrom's law, seeing as the channel volume shrinks going from 32nm (green) to 15nm (red) which would normally be associated with an increase in RDF and subsequent performance variation.

To explain the apparent reduction in JL-FinFET variability at finer technology nodes, we should determine the effect of scaling each physical dimension, i.e. L_g, T_{fin}, and H_{fin}, independently

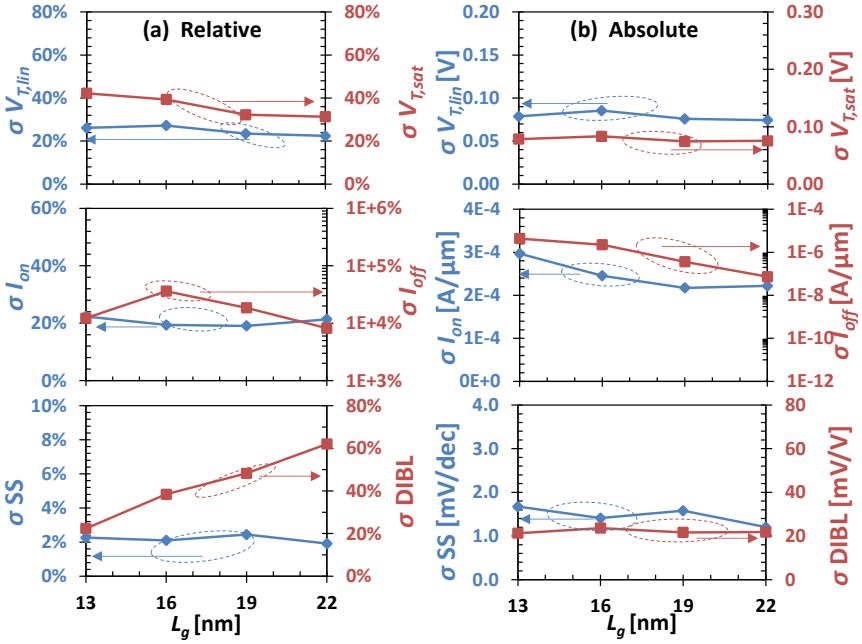


Fig. 17. (a) Relative variation and (b) absolute variation of JL-FinFET performance due to L_g scaling from 22 nm to 13 nm with H_{fin} = 10 nm.

of one another in order to isolate the cause of this behavior. The effect of scaling H_{fin} alone is already captured in Fig. 16 so we only need to focus on L_g and T_{fin}.

In Fig. 17(a), we show that when only L_g is scaled, the relative percentage variations do not follow any discernible patterns, despite our expectation that "Pelgrom-like" inverse relationships would appear. In other words, the trends cannot explain the observation in question. The sole exception appears to be σDIBL which appears to vary more at longer gate lengths and in reality is simply due to the baseline DIBL becoming progressively smaller at higher L_g values, while the non-normalized absolute variation remains constant as seen in Fig. 17(b). For the most part, the effect of scaling L_g has little impact on both relative and absolute variations of most parameters, and none of the trends appear to be responsible for the observation in question from Fig. 16.

On the other hand, when T_{fin} is scaled instead of L_g in Fig. 18 we see that smaller fin thicknesses result in less relative and absolute threshold voltage variation from RDF, which coincides

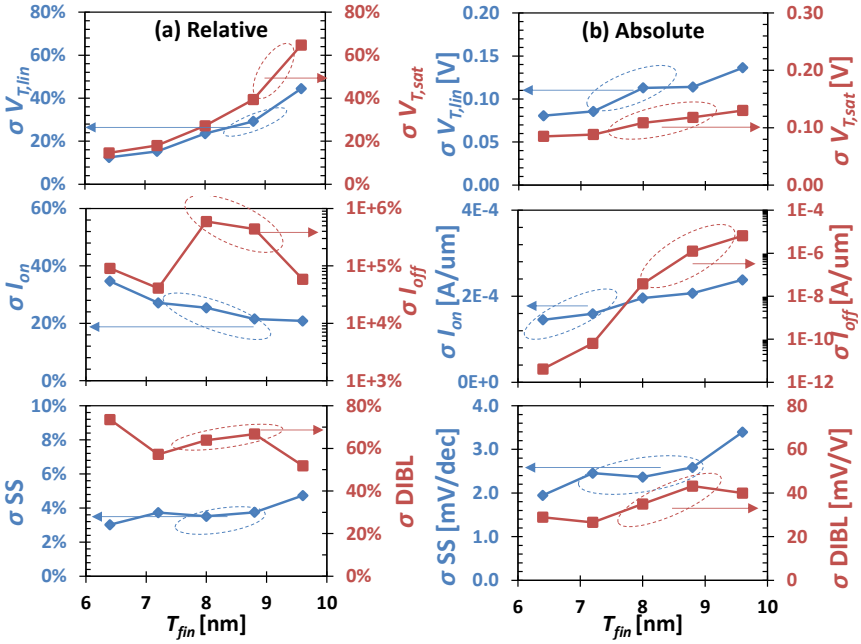


Fig. 18. (a) Relative variation and (b) absolute variation of JL-FinFET performance due to T_{fin} scaling from 9.6 to 6.4 nm with H_{fin} = 10 nm.

with the trend in question from Fig. 16. In addition, the magnitudes of $\sigma V_{T,lin}$ and $\sigma V_{T,sat}$ vary dramatically when T_{fin} is scaled from 6.4 to 9.6 nm, whereas the effect of varying L_g from 13 to 22 nm in Fig. 17 yields little to no response in $\sigma V_{T,lin}$ and $\sigma V_{T,sat}$. Recent work [39] has illustrated how the threshold voltage of JL-NWFETs becomes less sensitive to changes in channel doping as the nanowire geometry is miniaturized—a consequence arising from improved electrostatic control of the buried channel from the gate. Findings in [40] also agree, although our results suggest a much larger change in RDF sensitivity should be expected when the body dimension(s) are scaled to smaller values.

For other parameters such as I_{on}, I_{off}, and DIBL, the effect of scaling T_{fin} yields a more noticeable response in their respective variations, both relative and absolute. The trend for absolute σI_{on} is similar to that for $\sigma V_{T,lin}$ and $\sigma V_{T,sat}$ which is sensible since I_{on} is proportional to V_T. However relative σI_{on} exhibits the opposite trend, but this is simply because the baseline I_{on} is much higher for thick fins with poor short channel effect control. Absolute σI_{off} increases greatly with T_{fin} for similar reasons since I_{off} is exponentially related to V_T. Relative σI_{off} appears non-monotonic but this is most likely because absolute σI_{off} contains enough statistical noise that, when normalized to percentage values, results in uncertainty of up to two orders of magnitude. Similarly σDIBL appears non-monotonic but is likely due to statistical noise and not indicative of any physical "transition" from one regime to another.

To summarize, the effect of T_{fin} scaling in the presence of RDF is more significant than that of L_g scaling and its trend with σV_T versus T_{fin} may explain Fig. 16. When L_g and T_{fin} are scaled simultaneously, we may therefore expect the T_{fin} scaling behavior to dominate over that of L_g scaling to produce the overall technology scaling outcome.

Several caveats to our proposed explanation still remain, which we will now address. First, σI_{on} appears to be an exception to the results in that the 32, 21, and 15nm curves seem to overlap each other with no systematic scaling trend between them in Fig. 16. For example, since 32nm devices experience less RDF than 15nm devices, there may be an additional mechanism which compensates for this difference to ultimately yield a similar amount of σI_{on}. Knowing that σV_T is smaller (larger) for 15nm (32nm) devices, the mobility variation may be larger (smaller) due to the higher (lower) amount of RDF to produce comparable values of σI_{on} across all technologies. In other words, σV_T and mobility variation may oppose each other and cancel out in the determination of σI_{on}.

Second, we assume that scaling other parameters including the EOT, gate work function Ψ_M, and V_{DD} across different technology nodes change their baseline performance values but not their variations to any appreciable extent. We expect this assumption to be reasonable for Ψ_M and V_{DD} since they just shift the baseline V_T and limit the peak value of I_{on}, both of which are normalized out in the results of Fig. 16. Scaling the EOT likely has a minor influence on the variability data, but literature findings [41] suggest the effect to be small and in the opposite direction to Fig. 17(b). Thus we still expect T_{fin} scaling to be sole agent responsible for the trends in Fig. 16.

Overall, the variability impact of RDF is significant for JL-FETs and stems from the manner by which the intrinsic operation of depletion-mode transistors is affected. Recall from the previous section that LER potentially caused unwanted opening/closing of a conducting channel (Fig. 15), depending on the exact LER profiles along the fin sidewalls. In similar fashion, the exact number and positioning of ionized impurities can also result in fluctuation of the size and shape of the depletion region inside the fin. This, again, causes the buried channel to undulate with the topography of the randomized dopant profile (Fig. 13) and in some cases the buried channel may

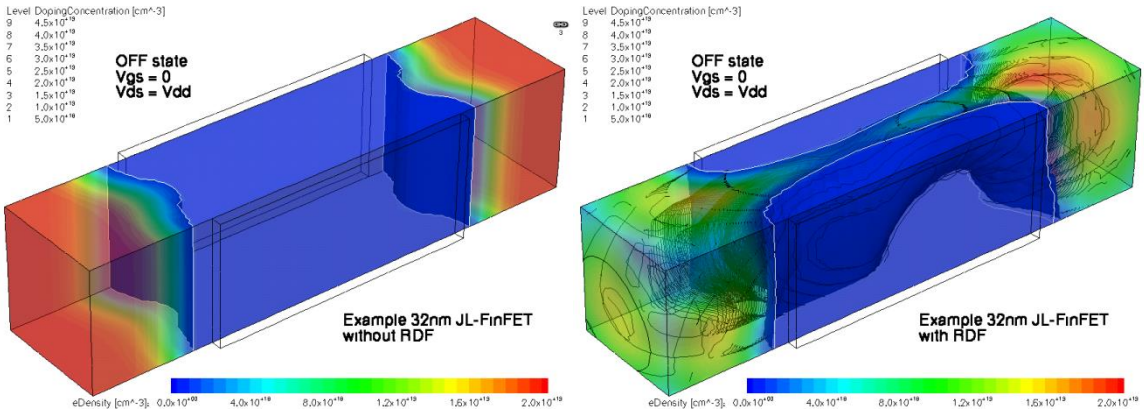


Fig. 19. Electron density plots in a representative 32nm JL-FinFET (H_{fin} = 10 nm) with and without RDF, showing the inadvertent formation of a conducting channel in the off state due to a surplus of dopants in the channel for the device with RDF.

be undesirably opened or closed as a result. For example, a JL-FET with too many dopants won't be fully depleted at zero gate bias and hence will remain "on" instead of "off" as shown in Fig. 19. Conversely, a JL-FET with too few dopants could remain fully depleted even at max gate bias and hence will remain "off" instead of "on". This is another characteristic weakness of depletion-mode FETs which rely on a delicate balance between electrostatic and geometric control for correct operation.

3.5 Circuit-Level Variability Impact[2]

Besides studying the impact of LER and RDF on the individual device level, it is important to consider how severe the circuit-level impact will be. In this section, we compare the impact of LER and RDF variability for IM- and JL-FinFETs for six transistor (6T) static random access memory (SRAM) cells as well as large-scale digital circuit benchmarks (i.e., microprocessors).

[2] This section describes work performed in collaboration with Shaodi Wang and Prof. Puneet Gupta from the UCLA NanoCAD Laboratory. I am very grateful for their many contributions to this section.

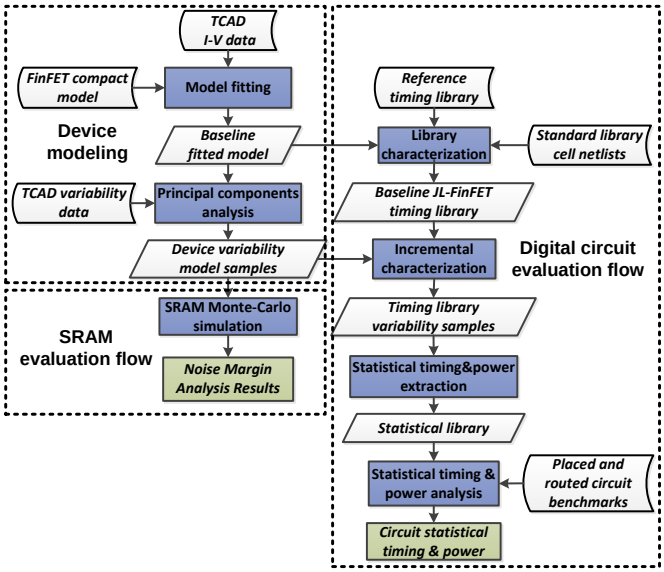


Fig. 20. Overview of the variability evaluation framework. The evaluation of 6T SRAM cells (left) and microprocessor circuits (right) are divided into two vertical branches as illustrated.

Our framework for evaluating circuit-level variability is represented in Fig. 20. Transistor *I-V* characteristics and variability data from the device-level TCAD simulations (Sections 2.4, 2.6, 3.3, and 3.4) are used as the starting input for subsequent compact modeling. In order to create a baseline model, we fit a BSIM model based on the Predictive Technology Model (PTM) [28] to match the TCAD I_D–V_G and I_D–V_D data. Using the method in Section 2.7 to capture the effect of LER/RDF in our compact model, model samples are generated such that their predicted behavior matches the original TCAD simulation results. 6T SRAM cell Monte Carlo simulations are performed by generating individual model samples for each of the six transistors, after which the static noise margin is extracted. For logic circuit timing and power analysis, we first create and characterize a baseline timing library from a baseline model and template library. Then, through incremental characterization based on model samples, library samples are generated such that the resulting circuit behavior should correctly reflect the performance impact from LER/RDF. Statistical timing and power information is extracted from these library samples, which are then fed as inputs to a computationally efficient statistical timing and power analyses tool based on [42], [43] to

Table 4 Allowed Tuning Range of Fitted Compact Model Parameters

Parameter	Range	Parameter	Range	Parameter	Range
nch	0.1-10x	len	0.7-1.6x	tox	0.7-1.6x
tsi	0.5-2x	tbox	0.5-2x	vth0(f)*	±0.25V
vth0(b)*	±0.25V	esi*	0.8-1.4x	eox*	0.8-1.4x
Lambda	0.5-2x	N*	0.9-1.1x	Vt*	±0.25V
voff1*	±0.1V	u0	0.7-1.6x	eta0	±0.1
dsub	±0.1V	rdsw	0.7-1.6x		

*Parameters in PTM model

evaluate the overall impact of LER/RDF on large-scale digital circuit delay and power consumption.

PTM FinFET models are fitted to the TCAD-simulated transfer and output characteristics. To match the currents from the 2-D TCAD simulations (in units of A/μm) to the 3-D device model, we linearly scale the currents to match single fin transistor characteristics, where we assume H_{fin} to be equal to the feature size in each technology node (e.g., H_{fin} = 32nm for 32nm FinFETs). Seventeen parameters of the PTM model are chosen as fitting variables according to the PTM and BSIM parameter extraction guide [28], [44] with tuning ranges for each chosen parameter listed in Table 4. Our error metric for the fitting procedure is the weighted least square difference between the simulated and model $I_D - V_{GS}$ and $I_D - V_{DS}$ curves, with random starts and gradient descent methods being applied. Good matching between the compact models against TCAD simulations are obtained, as illustrated in Fig. 21.

With the baseline compact model established, the baseline cell library is characterized using Nangate Open Cell Library [45] as the template. Extraction of device-level variability is based on principle component analysis (PCA) [31], [32]. The model samples are generated so that the resulting device performance variation matches the data from TCAD simulations. The statistical matching results are shown in Fig. 22. Standard deviations of I_{on} and $V_{T,sat}$ are calculated from 400 model samples. The maximum error is only 8.2% in σI_{on} for JL FinFETs, validating our JL FinFET circuit model. Unfortunately, when matching $\sigma V_{T,sat}$ for 15nm IM FinFETs, a maximum error of

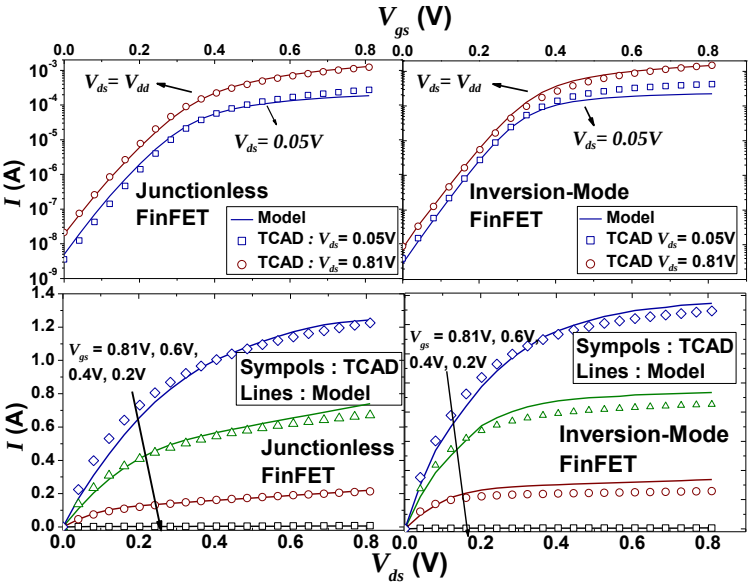


Fig. 21. Matching of baseline FinFET (a) transfer and (b) output curves between TCAD simulation and compact modeling.

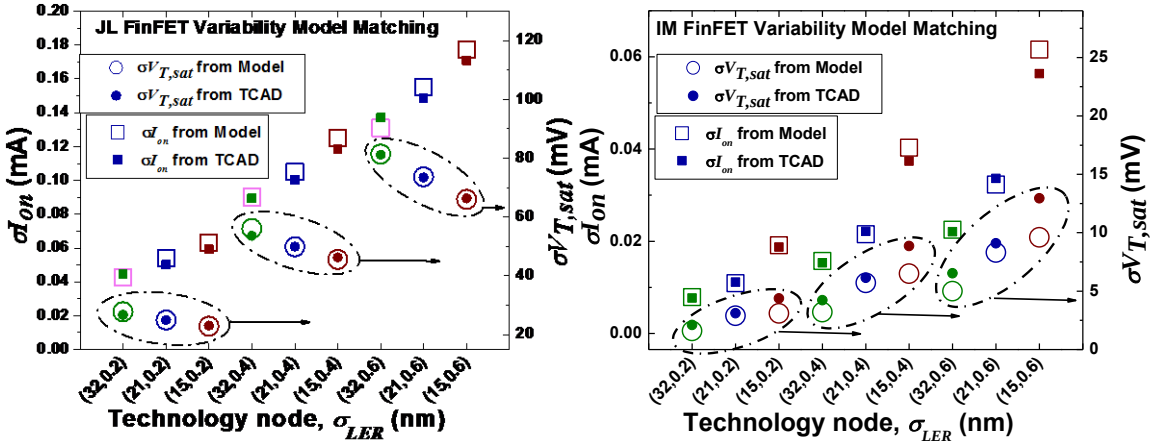


Fig. 22. Comparison of σI_{on} and $\sigma V_{T,sat}$ extracted from 200 samples between TCAD simulations and fitted variability models for (a) JL FinFETs and (b) IM FinFETs show a good fit.

25.8% is observed for σ_{LER} = 0.6 nm; however, since variation has very limited impact on IM FinFETs, we find that this relatively large matching error does not change our conclusions.

3.5.1 Variability Impact on 6T SRAM Cells

As CMOS technology continues to scale down, SRAM design becomes progressively more complicated. To guarantee proper operation, the cell design must meet noise margin requirements

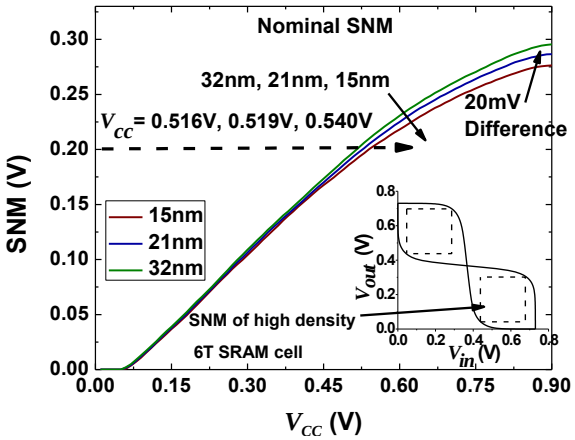


Fig. 23. Nominal SNM as a function of working V_{cc} for high density design JL FinFET 6T SRAM cells. Note that for successive technology nodes, SNM and $V_{cc,min}$ decrease when the other is held fixed.

which are budgeted for all fluctuation sources, including supply, process, and temperature variations. Increasing variability therefore strongly degrades performance. For instance, static noise margin (SNM), one of the important metrics for SRAM cell stability, decreases with successive technology generations [46]. Fig. 23 illustrates how nominal SNM changes with supply V_{cc} from 32nm to 15nm for JL FinFET 6T SRAM. With increasing V_{cc}, the SNM diverges for different technologies with differences of up to 20 mV at V_{cc} = 0.9 V.

In addition to these generic challenges, FinFETs face an additional disadvantage because of their digitized fin structures. Traditionally, device widths are sized to achieve high stability; for example, symmetric (SYM) designs might continuously scale PMOS widths to be larger size than NMOS to equalize the drive current. Realizing this with FinFETs requires parallelizing fins at the cost of cell area, for instance matching three PMOS with two NMOS fins; instead, typical designs now use one fin for each gate to maximize density [47], [48]. In the following discussion, all SRAM results are generated based on this high density (HD) layout unless otherwise specified.

3.5.1.1 Minimum Working V_{cc}

As cell density increases, power consumption becomes a crucial consideration requiring reduction of V_{cc} to conserve both dynamic and leakage power. The minimum working supply voltage V_{ccmin} is thus an important metric for judging the viability of a cell design. In general, for a fixed SNM, V_{ccmin} increases with scaling. Fig. 23 shows for instance how enforcing SNM of 0.2 V causes V_{ccmin} to increase from 0.516 V at the 32nm node to 0.540 V at 15nm. In addition to SNM, static/dynamic read and write noise margins also affect V_{ccmin}; however, considering all such metrics would raise many more design issues outside the scope of this paper. Therefore, we will only consider the effect of static noise margin on V_{ccmin}.

We use Monte Carlo simulations to search for V_{ccmin} under specified yield and SNM constraints. HSPICE is used for DC simulations of 6T SRAM cells where each individual device is independent and uses a randomly selected device model. The SNM is measured as the length of the largest square in the butterfly curve, as shown in the inset of Fig. 23. A simulated cell with SNM below the given constraint counts as a failed cell. A given supply voltage is said to work for SRAM cells if the number of successful simulations with this V_{cc} reaches the yield requirement (e.g., 99.9% yield requires 9,990 successful simulated cells out of 10,000 randomly generated cells). To find the V_{ccmin}, we use a binary search (40× faster than exhaustive search). To further improve runtime of yield analysis, we use the statistical blockade method [49] which uses rejection sampling, speeding up the total process by over 10×.

In Fig. 24, V_{ccmin} is reported for JL and IM SRAM cells with different technology nodes and LER amplitudes. The improved V_{ccmin} for IM-based SRAM compared to JL-based SRAM is explained by the fact that IM devices are more robust against LER-induced variability. This indi-

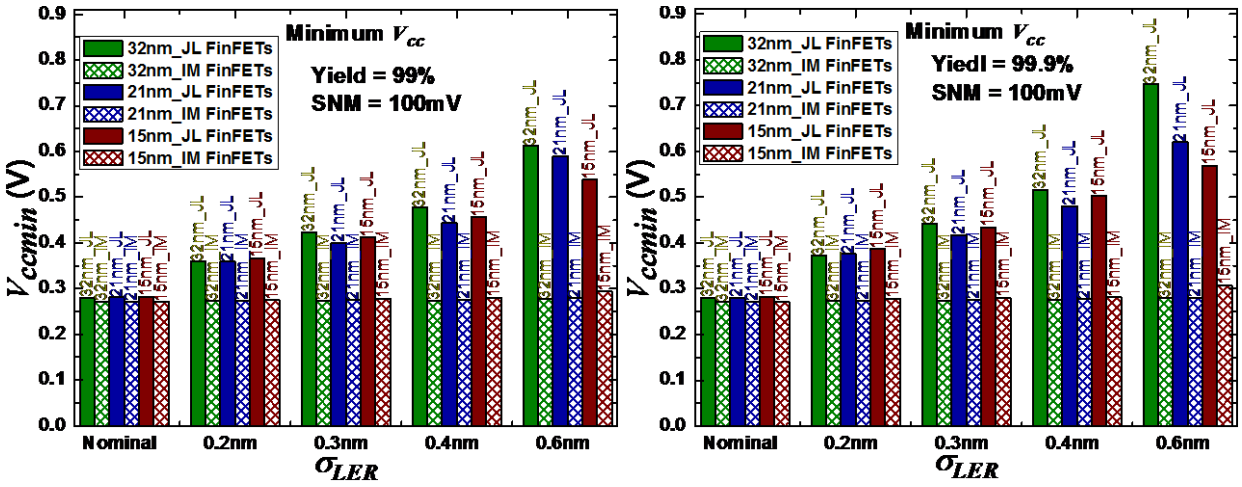


Fig. 24. V_{ccmin} as a function of technology node and LER amplitude for JL and IM FinFET 6T SRAM. The SNM constraint is 100 mV with 99% yield in the left panel, and 10 mV with 99.9% yield in the right panel.

cates that JL transistors in current technology nodes would not be a good option for memory design. This trend can be understood by remembering that V_{ccmin} is dictated by both variability and the nominal SNM. We have already seen that nominal SNM degrades under size scaling and dominates the trends in Fig. 24 at small σ_{LER}, but JL devices also become less sensitive to variability as technology scales (Fig. 14), allowing the operating conditions to relax. Our largest considered σ_{LER} of 0.6 nm is in line with the ITRS-projected σ_{LER} requirements of 1, 0.8, and 0.5 nm for the 32, 21, and 15nm nodes respectively. Therefore our results hold out hope that for realistic variability levels, JL SRAM technologies will become more competitive if scaling trends continue.

3.5.1.2 Static Noise Margin vs. Technology

We have also explored symmetric (SYM) SRAM designs using three PMOS with two NMOS fins, which can optimize nominal SNM and mitigate the effects of variability due to statistical averaging over the multiple fins. To characterize the impact of variability on the design, we define SNM loss as the percentage difference between the nominal SNM and the variability-

Table 5. Nominal SNM and SNM Loss from Variability for JL-FinFET Technologies

	32nm		21nm		15nm	
	HD[1]	SYM[2]	HD	SYM	HD	SYM
Nominal SNM[3] [V]	0.264	0.268	0.260	0.262	0.251	0.252
SNM w/ variation[4] [V]	0.128	0.154	0.144	0.166	0.140	0.176
% SNM loss	51.5%	42.5%	44.6%	36.6%	44.2%	30.2%

[1] High density 6T SRAM design
[2] Symmetric N/P design
[3] SNM at V_{cc} = 0.73V
[4] SNM with 99% yield constraint; LER variation (σ_{LER} = 0.6nm) @ V_{cc} = 0.73V

affected SNM. Table 5 compares SNM loss for junctionless HD and SYM cells. We find, as expected, that under scaling and/or use of SYM designs, SNM loss is significantly reduced. On the other hand, the symmetric design sacrifices read noise margin and cell area.

To better understand the impact of process variability on JL-FinFETs, we also attempted to incorporate both RDF and LER effects in our simulations, assuming the fluctuations to be uncorrelated. This assumption of statistical independence may not be strictly justified, but forms a best-case scenario for real-world situations. Even under this relaxed assumption, we find that no realistic V_{ccmin} can be realized for 99% yield and 100 mV SNM, reinforcing our conclusion that process variations will be a serious roadblock for JL FinFETs in memory applications.

3.5.2 LER Impact on Logic Circuit Variability

Although variability in JL FinFETs has a large impact at the device and cell level, large-scale circuits can mitigate and average out uncorrelated fluctuations. Analyses using closed-form analytical equations have shown how the number of gates and paths can decrease the overall circuit timing and power variations for conventional CMOS technologies [50]–[52]. Here we extend our methodology to analyze the usage of JL devices at the microprocessor level.

A typical way to analyze the statistical timing and power of circuit benchmarks uses a large number of library samples based on the Monte Carlo method (Section 2.7). However, this method is time-consuming and results in round-off errors when synthesizing tool outputs, losing statistical information. To fix these errors, more simulations are needed, with the quantity dependent on the

Table 6. Circuit Benchmarks

Technology Node	Frequency for Cortex-M0 [GHz]			Frequency for MIPS [GHz]		
	Fast	Typical	Slow	Fast	Typical	Slow
32nm	0.92	0.79	0.70	1.02	0.79	0.75
21nm	1.47	1.30	1.12	1.61	1.44	1.09
15nm	2.29	2.23	1.85	3.29	3.07	2.04

size of the variability impact. In this work, we use block-based statistical timing and leakage analysis [42], [43] to complete this step, drastically improving computational efficiency; in some cases, simulations which would previously require weeks of computation can be reduced to several tens of seconds.

To build the input to the statistical timer, the timing and leakage standard deviation for cells need to be extracted from library samples (we use 200 library samples in this step). We observe that timing variation is highly sensitive to input slew and output load capacitance. Hence, to find accurate timing variation information, a cubic model of delay standard deviation as a function of load capacitance and input slew is fitted to statistical timing information extracted from library samples. This model is found to be accurate enough for the following analyses. Leakage variation is modeled as a lognormal distribution with the standard deviation and mean extracted from the library samples.

The input to the statistical timer includes extracted timing models, extracted leakage lognormal standard deviations, a synthesized and routed circuit benchmark, the baseline library, timing constraints, and SPEF file containing parasitic information. For our benchmarking we select two processors, MIPS and Cortex-M0. To cover all working applications, we synthesize them in three operating clock frequencies for fast, typical, and slow speeds as shown in Table 6.

3.5.2.1 Circuit Simulation Results

Fig. 25 shows our results for MIPS designs. The clock period increase due to device variability is calculated as the sum of mean shift and delay uncertainty ($3\sigma_{clock}$), covering around 99.9%

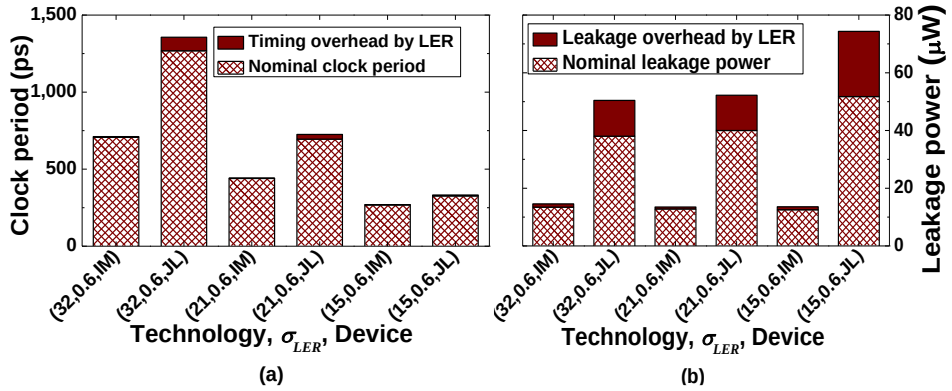


Fig. 25. (a) Nominal clock period and clock period increase (mean shift and variation) and (b) nominal leakage power and leakage power increase (mean shift and variation) due to LER variation (σ_{LER} = 0.6 nm) for IM and JL-FinFET-based MIPS processors at typical clock speeds.

of the possible clock period cases. All uncertainty in our timing results is below 1.20% of nominal delay. The mean clock period shift contributes the most; the highest mean shift is 7.04%. Thus, a delay margin of up to 8.2% may be needed to guarantee sufficient yield in the presence of LER. JL-based processors show a greater improvement in nominal speed with scaling compared to IM-based circuits.

The leakage power is assumed to follow a lognormal distribution. The uncertainty is calculated based on [43] at 99.9% yield point of leakage cases. Leakage increase is the sum of the mean shift and leakage uncertainty. As illustrated in Fig. 25(b), leakage power is severely impacted by LER. Our results show the increase mainly comes from a mean shift, in which the highest observed shift value is 43.02% of the nominal leakage. Leakage uncertainty has a considerable impact, inducing up to 15.57% increase. However, we expect that the leakage uncertainty will be negligible in industrial-scale designs (random leakage variation averages over number of devices in the design). High leakage variations are also predicted by device level simulations, where σI_{off} is over 10× nominal leakage for individual JL-FinFETs.

Fig. 26 and Fig. 27 show the JL-based high speed Cortex-M0 results for clock period mean and leakage mean compared with IM-based processors. JL devices are more severely affected by

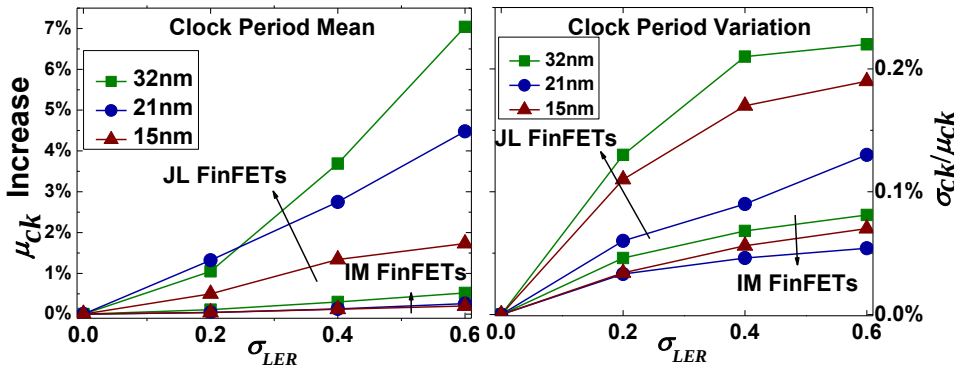


Fig. 26. (a) Increase in clock period mean and (b) variation of critical clock period as a function of technology node and LER amplitude for JL- & IM-FinFET circuit benchmark (Cortex-M0).

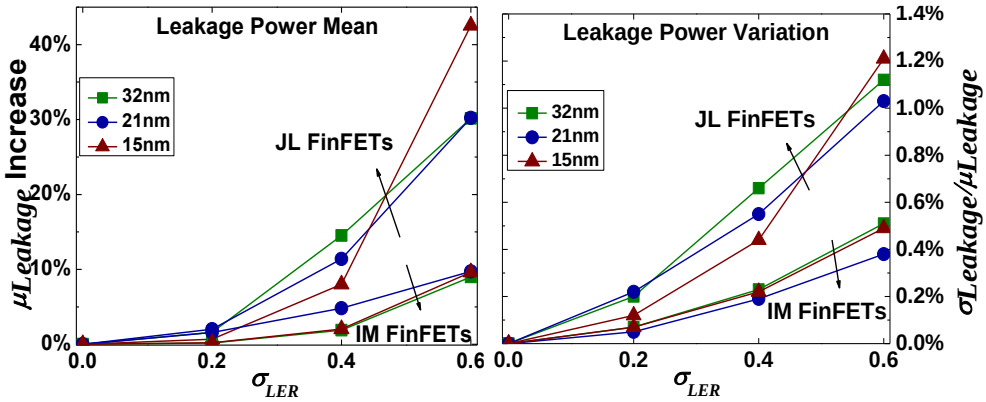


Fig. 27. (a) Increase in leakage power mean and (b) variation of leakage power as a function of technology node and LER amplitude for JL- & IM-FinFET circuit benchmarks (Cortex-M0).

variability in terms of both mean shift and standard deviation, with circuit clock period mean shift over 10× that of IM-FinFETs. Table 7 shows the average results from all six circuit benchmarks. For example, at σ_{LER} = 0.6 nm (near the ITRS predicted LER requirement of 0.5 nm), a 36.8% leakage mean increase is observed at the 15nm node. However, these impacts are not severe at the logic circuit level.

We have simulated the combined effects of RDF and LER variability, but the huge variations encountered (e.g., normalized $\sigma V_{T,sat}$ = 70%) can lead to statistically significant failure rates in SPICE convergence. Therefore these results are not presented. However, as previously observed [50]–[52], the mean increase of timing variations for circuits is linearly related to the variation of a single logic gate. We can estimate the combined variability to have 3× impact on

Table 7. Mean Shift and Standard Deviation of Timing and Leakage for Six Benchmark Circuits

Node	σ_{LER} [nm]	Timing		Leakage	
		μ_{delay}	σ_{delay}	$\mu_{leakage}$	$\sigma_{leakage}$
32nm	0.2	1.01%	0.12%	1.4%	0.2%
	0.4	2.56%	0.17%	12.6%	0.6%
	0.6	4.44%	0.22%	26.2%	1.0%
21nm	0.2	1.26%	0.13%	1.7%	0.2%
	0.4	2.30%	0.20%	9.6%	0.5%
	0.6	3.62%	0.27%	25.3%	0.9%
15nm	0.2	0.70%	0.17%	0.6%	0.1%
	0.4	1.32%	0.25%	6.8%	0.4%
	0.6	1.60%	0.28%	36.8%	1.1%

timing compared with our results considering only LER. For leakage power, a model-based analysis [43] using our library extraction results indicates the effects of combined variability will have 2× impact on leakage mean compared with the standalone LER variations.

3.6 Summary

The impact of LER on depletion-mode JL-FET (implemented as JL-FinFET) variability is dangerously high at the 32, 21, and 15nm nodes under current and projected lithography capabilities ($\sigma_{LER} \leq 1$ nm). Fluctuation in body thickness from LER results in direct modulation of the size and shape of the buried channel in JL-FETs, leading to unwanted opening/closing of a conducting channel which destroys proper switching functionality. Little distinction is found between the variability magnitudes between different technology generations; this fact is related to the primary agent responsible for LER-induced variability in JL-FETs. A maximum of 60% $\sigma V_{T,sat}$ is obtained for all JL-FinFET technologies at $\sigma_{LER} = 1$ nm which is 5× higher than for 15nm IM-FinFETs. σI_{on} also reaches up to 20% for all JL-FinFET technologies and is at least 3× higher than for IM-FinFETs. Leakage variation is exceptionally high with σI_{off} approaching 100,000% along with a hefty increase in mean I_{off} when LER is present. These results indicate that JL-FETs may have great

difficulty in meeting circuit requirements, especially those requiring precise matching of individual transistor characteristics. Circuit-level assessments indicate that variability in JL-FETs will be problematic for SRAM cells but remain manageable for large-scale benchmark circuits. While not explicitly investigated in this work, the adoption of spacer lithography will likely alleviate the LER burden to levels closer to (but larger than) spacer IM-FinFETs.

The impact of RDF on JL-FET variability is also dangerously high for the technology nodes under consideration. For minimum height devices (H_{fin} = 10 nm), performance variation from RDF is comparable to that from LER with 1 nm amplitude: $\sigma V_{T,sat}$ reaches up to 60% and σI_{on} up to 20%. Such high levels of RDF variability arise from the dependency of JL-FETs on the depletion region profile, which can undulate strongly due to random placement and number of dopant ions. Counter intuitively, we find that 32nm JL-FinFETs exhibit more RDF-induced variability than 15nm JL-FinFETs, despite having a larger active volume and dopant count. The results are explained by noting that the closer gate-to-channel proximity in thin body devices (15nm) reduces its sensitivity to RDF by means of stronger gate-channel coupling.

Both LER- and RDF-induced variability impacts are considered primary for JL-FETs and will likely pose significant challenges for the near-term adoption of JL-FET technology, unless significant improvements in lithography (i.e., reducing σ_{LER} or implementing spacer lithography) can be achieved. The key benefits of JL technology (e.g., manufacturing ease and scalability) must be weighed against the potential challenges associated with variability and lithography requirements to meet acceptable yields.

Chapter 4

Silicon Tunnel FET Variability

4.1 Background

TFETs [53]–[61] are promising devices for low power applications because of their potential to beat the 60 mV/dec subthreshold swing limit which has been a fundamental obstacle in today's MOSFET-based transistors. By using band-to-band tunneling (BTBT) instead of thermal diffusion to inject carriers from source-to-drain in a TFET, the minimum subthreshold swing is no longer tied to the Boltzmann-limited rate given by $kT/q \times \ln(10) \cong 60$ mV/dec at $T = 300$ K. Whereas MOSFETs are doped with the same polarity in both the source and drain (and channel for JL-FETs), TFETs are doped with opposite polarity in the source and drain with the channel usually left intrinsic, yielding a gated *p*-*i*-*n* structure. An example of an *n*-type silicon TFET is shown in Fig. 28 with a heavily doped *p*-type source, intrinsic channel, and *n*-type drain. In the "on" state, a high gate-source voltage causes enough band bending to facilitate BTBT of valence electrons in the source to the conduction band in the channel. In the "off" state, there is insufficient gate-source voltage which prevents proper band alignment to allow BTBT to occur.

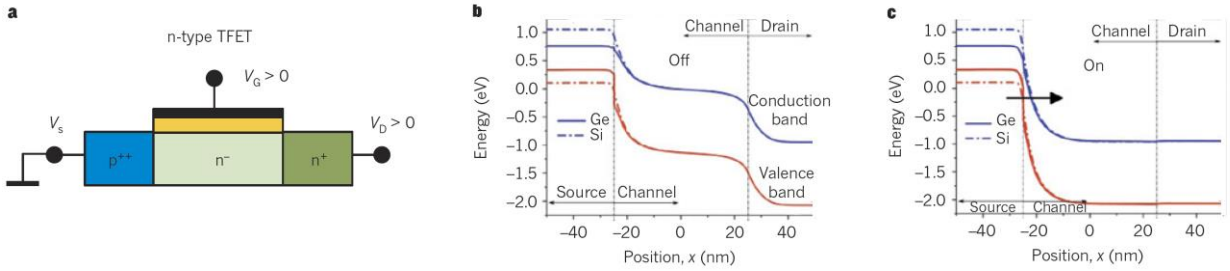


Fig. 28. (a) Structure of an *n*-type silicon TFET with *p*-type source, intrinsic channel, and *n*-type drain. (b) Band diagrams in the "off" state for all-silicon TFET and a silicon TFET with a Ge source. (c) Respective band diagrams in the "on" state. From [62]

Silicon TFETs are typically able to deliver very low off-state leakage currents on the order of fA/μm whereas MOSFETs are often limited to nA/μm of leakage current. Unfortunately, TFETs normally struggle to provide high on-state drive currents and typically max out in the range of μA/μm whereas silicon MOSFETs can easily deliver up to mA/μm of drive current. To improve I_{on}, heterostructures with Si/SiGe or Group III-V materials can be used at the source-channel junction to reduce the tunnel barrier [56]–[61]. Incorporating multi-gate configurations such as the DG or TG structure, or even nanowire gate-all-around (GAA) structures can also increase TFET performance by enhancing the gate-to-channel coupling to better modulate the BTBT current.

4.2 TFET Modeling

The TFETs considered in this study (example in Fig. 29) are modeled after the same *n*-type DG FinFET structures used in the previous chapters, with the exception of a *p*-*i*-*n* doping strategy in the source, channel, and drain regions, respectively. In this chapter, however, the simulated TFETs will not be exactly comparable in design to the IM- and JL-FinFETs targeted to meet the 2009 ITRS 32, 21, and 15nm high-performance logic nodes as before. This is due to the fact that no current ITRS guideline exists for TFETs, making a direct comparison against realistic FinFET technologies impossible. Instead, we will consider two hypothetical designs which utilize similar geometry and supply voltage values to the 32nm node, as will be described next.

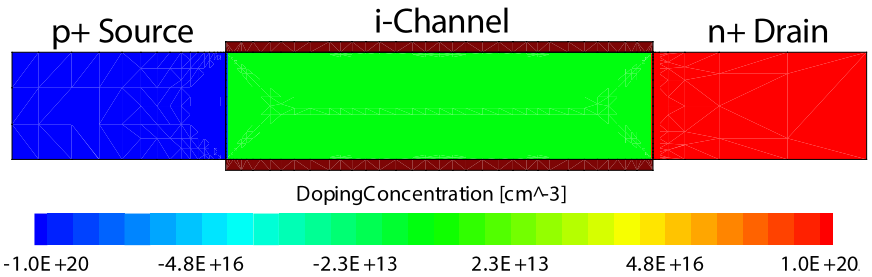


Fig. 29. Simulated *n*-type silicon DG TFET structure along with the doping strategy used in this work.

Table 8. Nominal Parameters for Simulated TFETs

Quantity	Technology Node		Description
	20/5	20/10	
L_g (nm)	20	20	Physical gate length
EOT (nm)	0.5	0.5	Equivalent oxide thickness
N (cm^{-3})	1×10^{20}	1×10^{20}	Source/drain doping
T_{body} (nm)	5	10	Body thickness
Ψ_M (eV)	4.15	4.15	Gate work function
V_{DD} (V)	1.0	1.0	Supply voltage
V_T (mV)	440	553	Threshold voltage
I_{on} (μA/μm)	0.227	0.171	On-state drive current
I_{off} (fA/μm)	0.190	0.134	Off-state leakage current
SS (mV/dec)	65.6	80.5	Subthreshold swing

V_T extracted at I_D = 1 nA/μm × W with $V_{DS} = V_{GS} = V_{DD}$.
I_{on} extracted at $V_{DS} = V_{GS} = V_{DD}$.
I_{off} extracted at $V_{DS} = V_{DD}$ and $V_{GS} = 0$.
SS calculated as $V_T / (\log 10^{-9} - \log I_{off})$

Specific parameters relating to the design of our TFETs are provided in the upper portion of Table 8. In this study, two designs are considered: the first (named "20/5") has a gate length L_g = 20 nm and body thickness T_{body} = 5 nm while the other (named "20/10") is designed with L_g = 20 nm and T_{body} = 10 nm. In both designs the source and drain regions are doped to a uniform N = 10^{20} cm^{-3} while the channel is left intrinsic, and we assume perfectly abrupt transitions between the source-channel and drain-channel junctions. For simplicity, no gate-source or gate-drain overlaps are present in our structures. Because of the symmetric doping for the source and drain, ambipolar behavior results and p-type conduction occurs for negative V_{GS} values. To avoid this issue, the gate work function is set to Ψ_M = 4.15 eV so that the minimum attainable current (i.e., I_{off}) occurs at V_{GS} = 0 and negative voltages are assumed to be unavailable. The equivalent oxide thickness (EOT) value of 0.5 nm implicitly represents that of a high-κ dielectric and we assume no gate leakage exists. Nominal performance values for the 20/5 and 20/10 TFETs are included in the lower portion of Table 8 showing threshold voltage V_T, I_{on}, I_{off}, and average SS.

A dynamic nonlocal BTBT model in Sentaurus based on Kane's model [63] is used to calculate the tunneling probability in different directions and the associated electron-hole generation rates at the starting and ending position along each path. The A and B coefficients in Kane's

formulation are calibrated to match experimental bulk *p-i-n* diode tunneling data from [64] and *p-n* junction data from [65]. The calibrated values used in our study are $A = 5.2\times10^{15}$ cm^{-3} s^{-1} and $B = 2.3\times10^{7}$ V cm^{-1}. This strategy was adopted to ensure that the basic tunneling process is easily and accurately captured with the BTBT model used here since directly equivalent all-silicon experimental TFET data is unavailable and also highly dependent on the specific implementation of experimental devices.

While the nonlocal BTBT model properly captures the on-state behavior of TFETs, treatment of the off-state regime requires additional modeling especially if trap-assisted tunneling becomes the dominant leakage mechanism instead of reverse *p-i-n* diode leakage. Thus, the actual leakage floor in a nanoscale TFET can be larger than what would be predicted by drift-diffusion models alone. In our simulations, we find that activating field-assisted Shockley Read-Hall (SRH) recombination models significantly worsens convergence, so instead we solve for the $I_D - V_{GS}$ (with $V_{DS} = V_{DD}$) curve in two stages. First, the $V_{GS} = 0$ point is solved using the Hurkx model with doping dependent [66] and field-assisted [67] lifetimes for electrons and holes in addition to BTBT in order to determine I_{off}. In this case, the trap states correspond to the (mid-gap) gold acceptor level. Second, the full $I_D - V_{GS}$ curve is solved using only electron BTBT to determine V_T and I_{on}. Band gap narrowing is accounted for by the Slotboom model and quantum corrections by the modified local density approximation (MLDA) [68].

Before proceeding, we must state a few disclaimers about the quantum correction model used here. For the body dimensions considered in our study (e.g. T_{body} = 5 nm to 10 nm), strong quantization in carrier densities should be expected, as well as subband splitting leading to effective band gap values larger than that of bulk silicon (i.e. $\Delta E_g > 0$). These effects and their impact on BTBT current can only be accurately captured via full quantum transport simulations [59] [70]

which are, unfortunately, computationally prohibitive in a large-scale study like this. The MLDA, while easy to implement, does not actually model band structure changes and does not interact with the Kane model; this is also similar to the DGA, however the DGA failed to converge in our simulations, leaving the MLDA as our only option. Furthermore, the Kane model parameters (*A* and *B*) are calibrated against bulk diodes and may become a function of T_{body} in quantized TFETs. Since we are mainly concerned with relative variability trends, fully quantitative accuracy is not required, and these models are assumed to be sufficient for our purposes. To verify this, we will include data obtained using the 1-D Schrodinger model to extract ΔE_g for each T_{body} and augment the BTBT parameters accordingly for comparison with the MLDA in the next section.

Finally, LER and RDF are modeled in the same manner as they were for IM- and JL-FinFET analysis, where we consider σ_{LER} up to 1 nm and treat all edges as uncorrelated, and dopant profiles are randomized using the Sano method [25], [69]. Example TFETs with and without LER and RDF are shown in Fig. 30. When discussing TFETs, we will use the term "body LER" instead of "fin LER" (or simply "LER"). For RDF, two major consequences arise from the non-uniform doping profiles: 1) the junction abruptness is "smeared out which can significantly impact the BTBT current; and 2) the effective channel length tends to shrink due to the smeared junctions. The latter observation is largely responsible for RDF-induced variability in planar MOSFETs [17]

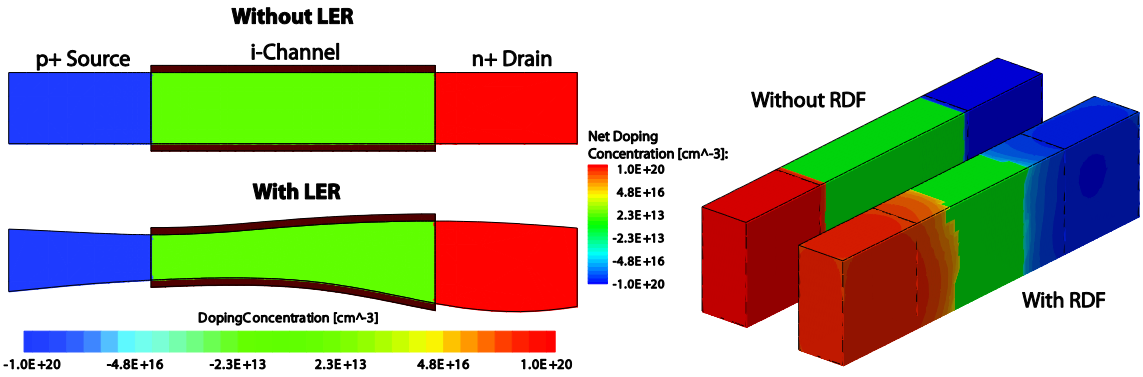


Fig. 30. Examples of simulated 20/5 TFETs with and without LER and RDF.

and IM-FinFETs (Chapter III), while the former will likely have major ramifications for TFET variability, as we will soon discover.

4.3 Baseline TFET Scaling

Before delving into the variability results, it is worthwhile to scale the nominal T_{body} parameter in the ideal TFET structure (without LER or RDF) and examine the subsequent performance impact in terms of the baseline V_T, I_{on}, I_{off}, and SS values. These results will become useful later on, and we will highlight any notable similarities or differences between our findings and those presented in [71] to which our side study is comparable to.

The baseline scaling trends are shown in Fig. 31 where the aforementioned performance figures are plotted versus T_{body} as the body thickness is scaled from 3 nm to 12 nm. From the trends

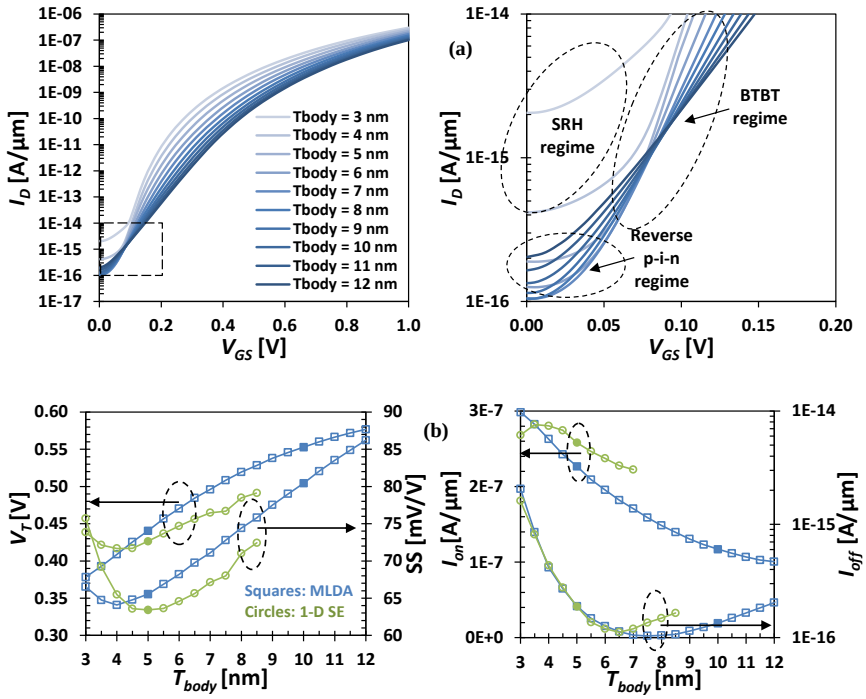


Fig. 31. (a) Raw $I_D - V_{GS}$ curves for the ideal TFETs with V_{GS} swept from 0 to V_{DD} and $V_{DS} = V_{DD}$. (b) Nominal TFET performance versus body thickness scaling from 3 nm to 12 nm in terms of the metrics V_T, I_{on}, I_{off}, and SS. Solid markers indicate the performance of 20/5 and 20/10 TFETs. Curves for the 1-D Schrodinger model are incomplete due to convergence problems.

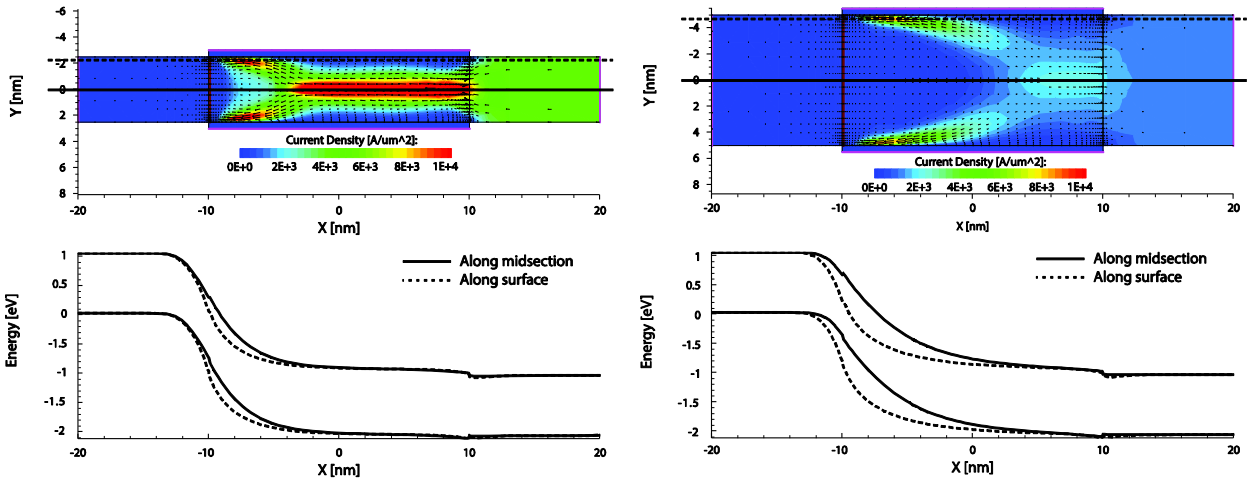


Fig. 32. On-state ($V_{GS} = V_{DS} = V_{DD}$) current density maps for the ideal 20/5 TFET (left) and 20/10 TFET (right), along with the energy band diagrams along two horizontal cut lines: one along the body midsection (solid) and another near the silicon-SiO_2 interface (dashed). In thin body TFETs, the energy bands are sufficiently lowered by the high gate voltage to induce BTBT along the midsection in addition to the two surface channels. In thick body TFETs, significant BTBT only occurs along the two surface channels and not along the midsection.

depicted with the MLDA activated, I_{on} is degraded as the body thickness increases; this is mainly due to the weaker combined electrostatic control of the channel midsection from both gates, thereby resulting in less band bending at the source-channel junction and a longer tunneling path (Fig. 32) near the midsection. In general, this is also true for the tunneling current at any V_{GS} value where BTBT is important as revealed by the raw $I_D - V_{GS}$ curves in Fig. 31(a). This contradicts an observation found in [71] where the drive current actually shrinks towards very small T_{body} values (below 7 nm); however, no concrete explanation was given besides a possible reduction in cross-sectional area for current flow. In reality, quantization can lead to effectively larger band gaps for very thin bodies and result in less BTBT current, causing a reduction in drive current. This is confirmed when the 1-D Schrodinger model is activated in lieu of the MLDA as shown in Fig. 31(b). For $T_{body} > 5$ nm, however, similar trends are produced using either model.

Furthermore, since V_T represents the gate voltage required to reach a certain current, the threshold voltage gets larger as T_{body} increases and is also a consequence of the weaker electrostatic

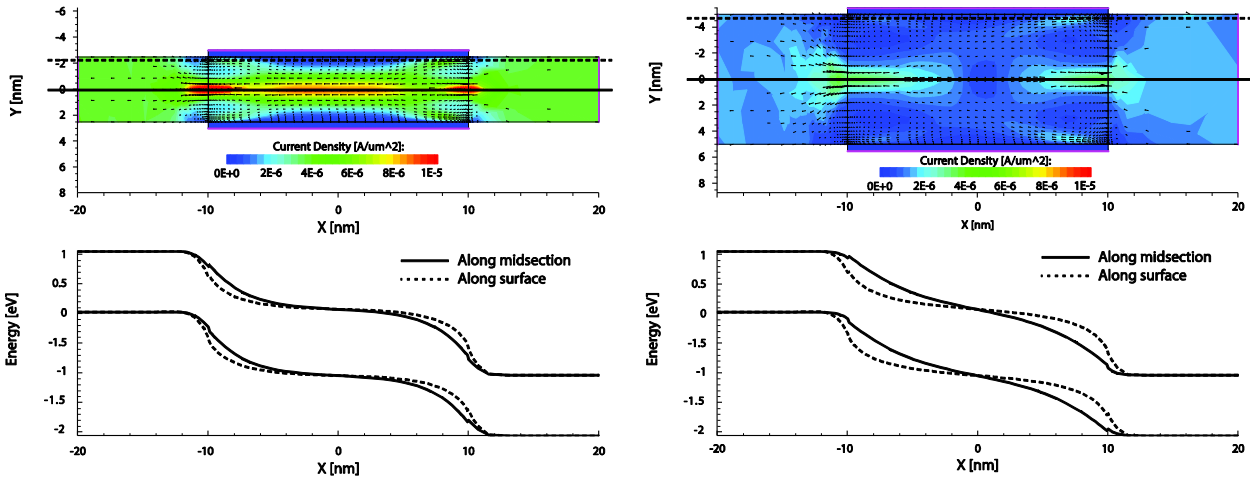


Fig. 33. Off-state (V_{GS} = 0 and V_{DS} = V_{DD}) current density maps for the ideal 20/5 TFET (left) and 20/10 TFET (right), along with the energy band diagrams along two horizontal cut lines: one along the body midsection (solid) and another near the silicon-SiO_2 interface (dashed). In thin body TFETs, the barrier height for trap-assisted tunneling is lowered by the close proximity of the gate to all vertical locations in the channel compared to thick body TFETs where the gate loses control of the midsection, resulting in a larger barrier to prevent significant tunneling through traps.

gate control. While neither V_T nor I_{on} appear to scale linearly with the body thickness, their sensitivities (or derivatives with respect) to T_{body} centered at any point do appear linear over a range of ±2 to ±3 nm; this will become important later on.

The relationship between I_{off} with T_{body} is more complex, with a non-monotonic trend appearing in the lower right plot of Fig. 31(b). Comparing these I_{off} values with the low V_{GS} portions of Fig. 31(a), we see that standard reverse *p-i-n* leakage dominates for large body thicknesses (T_{body} > 7.5 nm) while SRH-induced leakage dominates for small body thicknesses (T_{body} < 7.5 nm). As the body thickness increases, the magnitude of reverse *p-i-n* leakage increases due to the larger cross-sectional area, while field-assisted tunneling through deep-level traps becomes less probable as the gate loses control of the body's midsection and cannot force enough band bending to reduce the tunneling path from the valence band to the trap level, and finally to the conduction band (Fig. 33). These findings are also different from those in [71] since the authors there did not consider SRH leakage mechanisms in their simulations. The SRH leakage regime is unchanged between

the MLDA and 1-D Schrodinger models because neither model includes a parameter to account for ΔE_g; we concede this as a modeling limitation and hence, I_{off} may be overestimated when T_{body} < 5 nm. Generally speaking, we find that the relationship between I_{off} and T_{body} is highly nonlinear within the *p-i-n* dominated and SRH dominated regimes, with characteristics resembling exponential-like behavior.

Finally, because the average subthreshold swing is determined by I_{off} and V_T according to the definition provided in Table 8, the relationship between SS and T_{body} can be deduced from the previous trends, namely I_{off} vs. T_{body} and V_T vs. T_{body}. Like the off-state current, the subthreshold swing exhibits a non-monotonic trend with body thickness; however, the minimum SS value occurs at a lower T_{body} value (= 4 nm for MLDA) compared to the T_{body} value (= 7.5 nm) for minimum I_{off}. This arises from the lower threshold voltage for thin bodies, which manages to reduce the subthreshold swing despite the sub-optimal leakage current value. Interestingly, the SS vs. T_{body} curve is quite linear for T_{body} > 4 nm.

These trends will aid us in explaining the LER results in the next subsection assuming the effect of LER, to first order, can be simplified as a fluctuation in the average (uniform) body thickness inside a TFET. Essentially, this represents a Taylor-series approximation, wherein the standard deviation in a parameter *P* due to LER can be approximated as

$$\sigma P = \frac{dP}{dT_{body}} \sigma_{LWR} = \sqrt{2} \frac{dP}{dT_{body}} \sigma_{LER} \tag{3}$$

where dP/dT_{body} is the sensitivity of the parameter *P* with respect to T_{body}, which can be obtained from the curve tangents in Fig. 31(b). Although this ignores the impact of non-uniformities in body thickness along the channel, as shown in Fig. 30, it provides us with some basic intuition for how LER can manifest itself in the overall variability of TFET performance.

We must remind the reader that the full influence of random spatial non-uniformities cannot be easily (or at all) modeled in closed form and can only be rigorously captured through brute force statistical TCAD simulations, which is why we still perform them in this study. In summary, the approach in (3) serves as an approximation at best and cannot provide accurate estimates of variability magnitudes, however it can provide useful insight to assess and explain the sensitivity of TFET technology to geometric variability sources such as LER.

4.4 LER-Induced TFET Variability

The performance variability of 20/5 and 20/10 TFETs versus σ_{LER} are displayed in Fig. 34 in terms of the percentage standard deviations (relative to their nominal values in Table 8) of each metric discussed earlier, namely σV_T, σI_{on}, σI_{off}, and σSS. As a reminder, only the MLDA is used

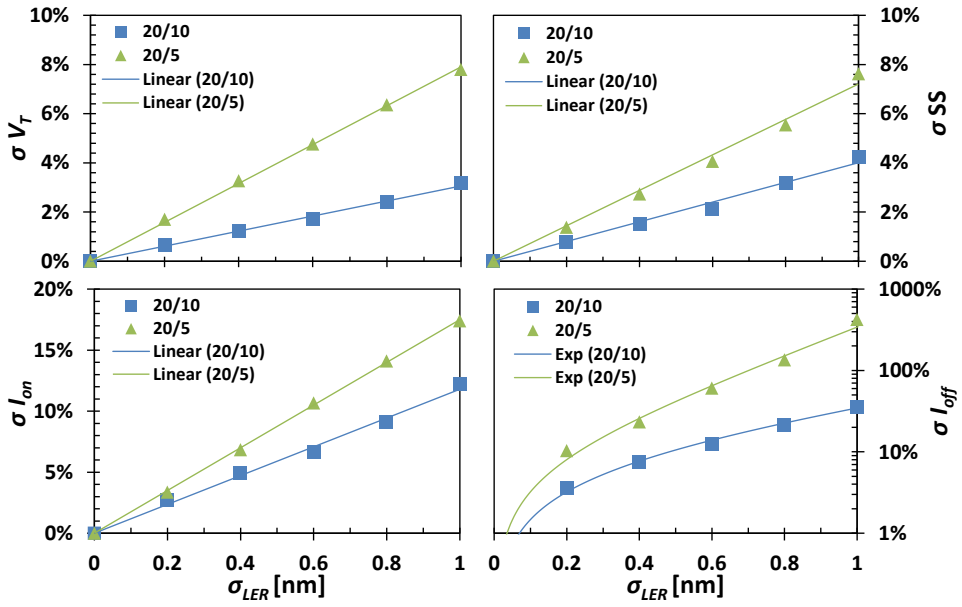


Fig. 34. Device-level variability of 20/5 and 20/10 TFETs due to body LER with σ_{LER} ranging from 0 to 1 nm. Markers indicate actual data while lines indicate best fits.

henceforth. Overall, the more aggressively scaled 20/5 TFET yields larger variation in every parameter compared to the 20/10 TFET—this is reasonable considering the larger fractional change in body thickness that LER causes for smaller devices.

We also observe that σV_T, σI_{on}, and σSS all exhibit very linear dependences with σ_{LER} (at least up to 1 nm). Invoking the simplified argument introduced in (3), the linear trends in Fig. 34 appear consistent with the observation that the nominal scaling behaviors in Fig. 31(b) are also linear over a range of at least ±1 nm centered at T_{body} = 5 nm and T_{body} = 10 nm for both TFET designs. On the other hand, σI_{off} exhibits an exponential dependence with σ_{LER} for both designs, yet this finding is also consistent with Fig. 31 where the off-state current in both the reverse saturation and SRH-dominated regimes scale exponentially with body thickness near T_{body} = 5 and 10 nm.

Overall, however, the magnitude of LER-induced variability for these TFETs can be considered manageable especially if we compare these fluctuations to those of standard double-gate FinFETs with roughly equivalent geometries (Fig. 5). We find that both σV_T and σSS are kept below 10% for the TFETs over the LER range studied which is comparable to the respective levels for standard IM-FinFETs. Unfortunately, we observe that drive current variation for TFETs is noticeably higher, with σI_{on} reaching between 10-20% at 1 nm LER, compared to only 3-6% for similarly sized IM-FinFETs at the same LER amplitude. In Fig. 35 we compare the relative change in I_{on} for a 20/5 TFET compared to an equivalent 20/5 IM-FinFET with identical LER patterns in both devices. Here, the TFET is more heavily impacted by the same LER pattern than the IM-FinFET; this is likely because I_{on} in an IM-FinFET is sensitive to the body thickness along the entire channel (since inversion occurs roughly along the entire channel), whereas in a TFET usually only the body thickness near the source-channel junction matters (since BTBT at a specific

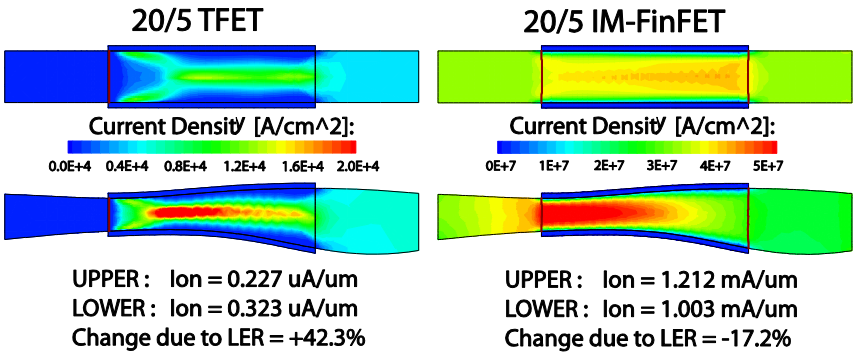


Fig. 35. Effect of a specific LER pattern (σ_{LER} = 1 nm) on the drive current of a 20/5 TFET compared to an equivalent IM-FinFET. Only the doping and work function are different between the two, all other parameters are identical.

location limits the channel resistance rather than inversion across the channel). In other words, the LER impact in a single IM-FinFET self-averages out over the entire channel as far as drive current is concerned, but it does not self-average out in a TFET due to the more localized nature of BTBT and hence, I_{on} is affected by body LER more than an IM-FinFET.

Finally, σI_{off} reaches up to 400% at 1 nm LER for the 20/5 design and is accompanied by a significant increase in mean leakage current (roughly 140%) due to the exponential dependence of I_{off} with T_{body}, which also produces a highly non-Gaussian probability distribution. A similar effect also occurs in FinFETs which results in exponential σI_{off} vs. σ_{LER} trends and log-normal I_{off} distributions as well (Fig. 5 and Fig. 14). For 20/10 TFETs, however, σI_{off} is reduced to only 35% at 1 nm LER but only with a sacrifice in drive current and subthreshold swing (Table 8). Comparing the σI_{off} and σI_{on} curves from Fig. 34, it is clear that leakage current variation is more sensitive to LER than drive current variation; this supports a similar conclusion in [70], despite very different modeling approaches. Because TFETs will most likely be targeted in the niche of low standby power circuits where their steep switching properties can be fully exploited, wider T_{body} designs should present a favorable tradeoff to achieve better control over leakage power at the cost of peak switching speed.

Based on the results presented, estimates of the maximum acceptable LER amplitude for a specific circuit application can be obtained either from the data in Fig. 34 directly, or by extrapolation of the fitted trend lines. We have investigated larger LER amplitudes up to σ_{LER} = 3 nm and find that σV_T, σI_{on}, and σSS more or less retain their linear dependencies with σ_{LER}, but with less stability toward higher σ_{LER} values; we attribute this to inability of the TCAD software to properly simulate structures which exhibit wildly irregular geometries due to the large LER present using the physical models in this work. In other words, we cannot trust the simulation models to accurately describe the behavior of some TFETs with extreme roughness—these instances are manifested as statistical outliers. Moreover, 20/5 TFETs with $\sigma_{LER} \gg 1$ nm have a high probability of structural failure or near-failure in which the channel becomes discontiguous (i.e. broken)—this would clearly be unacceptable from a manufacturing standpoint and presents an upper limit on the maximum permissible LER amplitude from a processing standpoint. Spacer lithography may help this situation by reducing σ_{LWR} for a given σ_{LER} via line edge correlation and produce less LER-induced variability for TFETs, like it does for FinFETs (Fig. 5). Meanwhile, our results for $\sigma_{LER} \leq$ 1 nm are sufficiently reliable to support our main conclusions.

4.5 RDF-Induced TFET Variability

The variability impact of RDF on our 20/5 and 20/10 TFETs is shown in Fig. 36 for different device heights H ranging from 10 to 40 nm, representing typical values that may be used in actual fabrication. Shorter device heights yield more RDF variability due to the fractionally larger variation in dopant population that occurs in smaller device volumes. This is consistent with Pelgrom's scaling law (2) for traditional MOSFETs, where H replaces the device channel width. Comparing the 20/5 and 20/10 curves, we observe that smaller body widths yield more variability

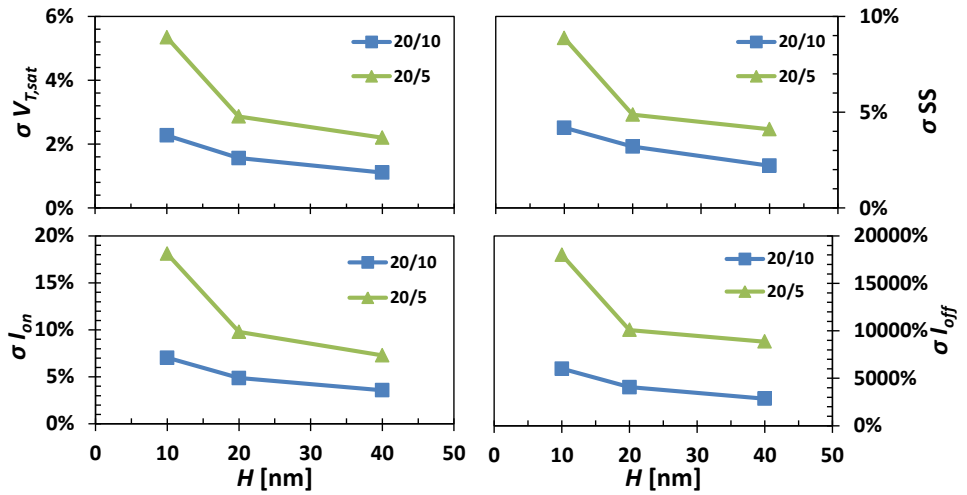


Fig. 36. Device-level variability of 20/5 and 20/10 TFETs due to RDF for different device heights H from 10 to 40 nm.

compared to larger body widths, which also follows from the previous argument. Comparing Fig. 36 with Fig. 34, the amount of RDF-induced variability is similar in magnitude to that due to body LER in the range of $\sigma_{LER} \leq 1$ nm; interestingly, a similar conclusion is reached for JL-FinFETs as well (Fig. 16).

Unfortunately, the RDF model adopted in this work predicts an additional problem: significant loss in nominal TFET performance arising from smeared junctions. This is highlighted in Table 9 and Fig. 37, which shows major degradation in average I_{on}, I_{off}, and SS for TFETs with RDF compared to their baseline values. Intuitively, this occurs because the source-channel junction is no longer perfectly abrupt, and consequently any increase/decrease in applied gate voltage will have a weaker impact on the peak electric field and tunneling barrier width. This results in

Table 9. Comparison of Average Versus Nominal TFET Performance With and Without RDF

Quantity	Without RDF (Nominal)		With RDF (Average)	
	20/5	20/10	20/5	20/10
V_T (mV)	440	553	535	651
I_{on} (μA/μm)	0.227	0.171	0.087	0.040
I_{off} (fA/μm)	0.190	0.134	227	130
SS (mV/dec)	65.6	80.5	146.7	167.5

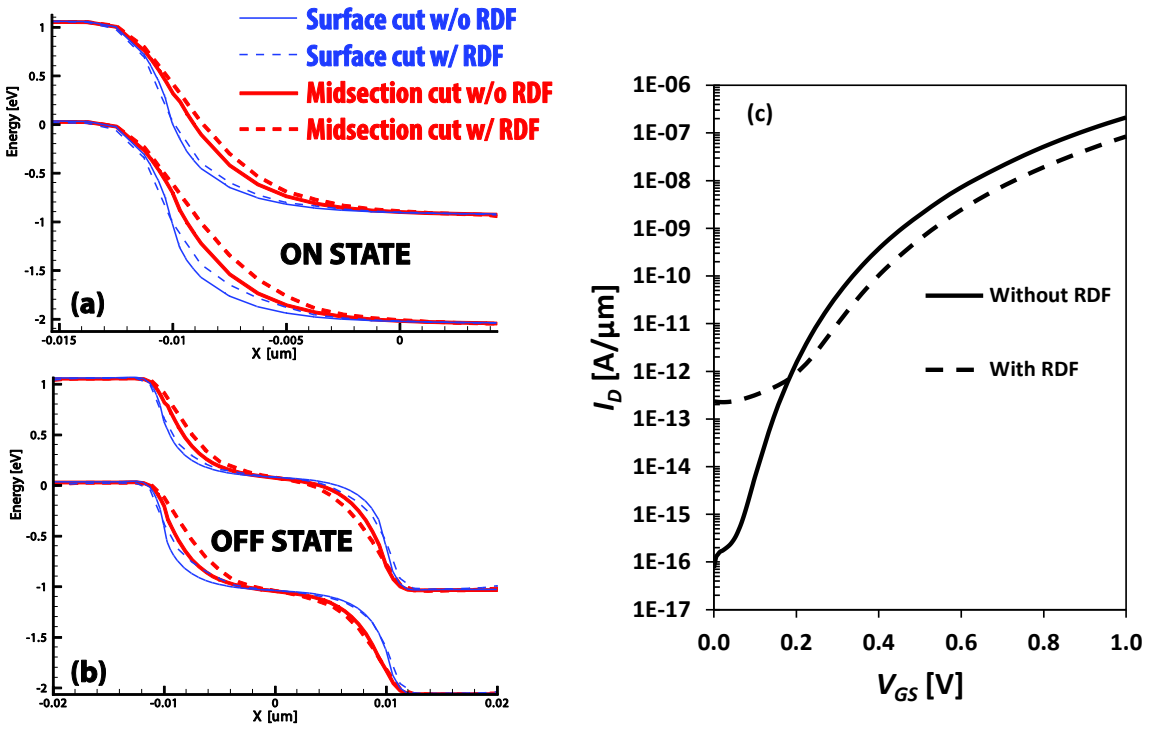


Fig. 37. (a) On-state and (b) off-state band diagrams for 20/5 TFETs with and without RDF along the channel surface and channel midsection. RDF causes the source-channel tunneling path to slightly widen in the on-state, while the direct source-to-drain trap-assisted tunneling path shortens in the off-state. (c) Corresponding $I_D - V_{GS}$ curves with and without RDF, showing severe degradation in performance predicted from the RDF model.

weaker gate control of the tunnel barrier, as more of the band bending is distributed laterally over the entire channel length rather than localized at the source junction, as shown in Fig. 37(a), resulting in less drive current. Leakage current is significantly worse because the direct source-to-drain (trap-assisted) tunneling path is shortened from the smeared junctions, as shown in Fig. 37(b). Both of these effects are captured by the degraded I_D - V_{GS} curves in Fig. 37(c).

While the mean shifts in performance metrics appear strong, their effects are masked by the presentation of relative variations (rather than absolute variations) in Fig. 36 when percentages are taken. Additionally, the long-range potential components from individual dopants may encroach into the intrinsic channel, especially if the dopants are located near the source-channel junction. This effect results in a shorter effective channel length compared to the physical L_g as

observed in Fig. 30. However, this is a secondary effect and largely not responsible for the behaviors in Table 9 or Fig. 36, as the actual channel length is still not very significant in determining TFET performance at these geometries. To verify this, we considered "40/10" TFETs and found the LER and RDF trends to be nearly identical to the 20/10 TFETs.

Nevertheless, junction smearing in light of RDF may pose a serious problem for targeting TFETs to meet aggressive I_{on} vs. I_{off} requirements. The reader should keep in mind, however, that these findings are contingent upon the validity of the RDF model when used with the nonlocal BTBT model. To our knowledge, there is no consensus on whether Sano's model is physically consistent with Kane's, despite being used individually with success in TCAD simulations. The authors of [69] made a similar disclaimer as well despite not mentioning the junction "smearing" effect. Clearly, this is a shortcoming that needs to be addressed in future literature and, until then, the results in this section should be approached with caution.

4.6 Summary

The impact of LER on TFET variability is moderate when compared to that on IM- and JL-FinFETs. Extracted performance variations are found to be linear for σV_T, σI_{on}, and σSS, and exponential for σI_{off} when plotted against σ_{LER} up to 1 nm, with thinner body TFETs (20/5) experiencing more variation than those with wider bodies (20/10). Up to 8% σV_T, 17% σI_{on}, and 300% σI_{off} are obtained at σ_{LER} = 1 nm for 20/5 TFETs; these variability magnitudes are mostly similar to those for 15nm IM-FinFETs at the same LER values, except for a higher σI_{on} in TFET case. The variability of TFETs from LER is somewhat worse than IM-FinFETs due to the localized presence of BTBT at the source-channel junction compared to the distributed presence of inversion carriers along the entire channel in a FinFET. As a result, nonuniformities in TFET body thickness do not

self-average across the channel length the way they do in FinFETs; this results in greater interdevice variability for TFETs compared to FinFETs. High sensitivity of the BTBT process to subtle changes in tunneling distance at the source-channel junction also contribute to the higher LER-induced variability in TFETs compared to IM-FinFETs. However, the surface-channel characteristic of TFETs ensures that LER does not have a primary effect on TFET operability, unlike the case for JL-FETs with buried channels. As a result, TFET variability due to LER remains much lower than for JL-FETs. While not explicitly investigated in this work, the adoption of spacer lithography may also reduce the LER impact by keeping T_{body} near the source-channel junction consistent between different TFETs.

The impact of RDF on TFET variability is also moderate, but with different implications for device design. Like the case for minimum height (H = 10 nm) FinFETs, TFET performance variation from RDF is, at first glance, comparable to that from LER with 1 nm amplitude: σV_T reaches up to 6% and σI_{on} reaches up to 18%. However, we find that σI_{off} from RDF (nearly 20,000%) is much larger than that from 1 nm LER (up to 300%). The high leakage current variation is accompanied by a large increase in mean I_{off} compared to the baseline value, and we attribute this to a junction "smearing" effect from the RDF model. Non-abruptness of the source-channel junction due to discrete dopants at the atomistic regime suggests that the simple assumption of abrupt junctions is highly unrealistic and will overestimate TFET performance in terms of I_{on}/I_{off}. Since engineering of the source-channel junction is a key design aspect in TFETs, the impact of RDF must not be ignored and may be particularly strong in deeply-scaled generations for low power applications where I_{off} becomes a critical metric. Nevertheless, TFET variability due to RDF remains a secondary effect because the intrinsic device operation is not directly jeopardized, unlike the case for JL-FETs.

Both LER- and RDF-induced variability impacts are considered secondary for TFETs and should not pose major obstacles to adoption of TFET technology, assuming proper care is taken during device design and optimization in light of junction non-abruptness. Future circuit-level variability investigations will be a valuable asset to further assess the viability of TFETs as a viable MOSFET replacement technology.

Chapter 5

Interactions between LER and RDF in Nonplanar FET Variability

5.1 Background

So far, we have seen that device variability from LER and RDF present obstacles to the scaling of both planar and non-planar CMOS technologies. In Chapter 2, we saw how multi-gate transistors such as FinFETs are particularly vulnerable to LER because of their heavy reliance on physical geometry to ensure good electrostatic integrity between the gate and channel regions. Meanwhile, devices containing heavily doped active regions, e.g., JL-FETs, are greatly affected by RDF when scaled to nanometer dimensions, as we saw in Chapter 3. Furthermore, TFETs which rely on effective modulation of an energy barrier within a highly abrupt source-channel junction also show sensitivity to LER and RDF as demonstrated in Chapter 4. Consensus holds that such variability mechanisms will play a major role in the feasibility of adopting next-generation FET technologies to replace silicon CMOS.

Until now, we have treated the effects of LER and RDF independently in order to gain insight into how each variability mechanism affects FET behavior. However, real devices are impacted by numerous sources of variation simultaneously, and these sources may or may not interact with one another. For planar CMOS, the statistical effects of LER and RDF are presumed to combine independently [72], allowing us to separately treat the impacts of (gate) LER and RDF in the transistor active regions and arrive at a simple picture which describes how each variability mechanism causes performance variation. This approach has been validated against simultaneous LER

and RDF simulations [72], confirming the independence of LER and RDF effects in planar technologies. However, it is uncertain whether this conclusion holds for non-planar technologies as well.

In this chapter, we investigate whether or not the effects of LER and RDF are independent in non-planar technologies such as those based on FinFET designs. Using TCAD simulations, we demonstrate that an accurate prediction of device variability from LER and RDF cannot be obtained for IM-FinFETs and TFETs by modeling LER and RDF separately from one another. On the other hand, we find that JL-FET variability due to LER and RDF is still reasonably well described even when the two mechanisms are separately modeled and combined in an uncorrelated fashion. These findings are explained based on the underlying structure and mode of operation for each FET type, and provide insight for meaningful evaluation of near-future variability scenarios.

5.2 Modeling Approach

Our technologies of interest are double-gate IM-FinFETs, JL-FETs, and TFETs designed for the 2009 ITRS high performance logic 32, 21, and 15nm nodes. These are implemented in Sentaurus TCAD using 2-D (for LER analysis) and 3-D (for RDF analysis) simulations[3] and are identical to those in Chapter 2 through Chapter 4 for consistency. Some of these technologies are no longer considered "near-future" (32 and 21nm) as of 2015; however, the results will not change our general conclusions which aim to compare different FET varieties, e.g. inversion-mode versus depletion-mode (junctionless) and thermal versus tunnel injection. Specific parameters related to the design and nominal performance of each device are summarized in Table 10. As mentioned in

[3] Slight differences in performance from the dissimilar meshes in LER and RDF simulations were found to be insignificant for comparison purposes here.

Table 10. Nominal Parameters for Simulated FETs

Quantity	Inversion-Mode FinFET			Junctionless FinFET			Tunnel FET		Description
	32nm	21nm	15nm	32nm	21nm	15nm	20/5	20/10	
L_g (nm)	22	17	13	22	17	13	20	20	Physical gate length
EOT (nm)	0.90	0.77	0.64	0.90	0.77	0.64	1	1	Equivalent oxide thickness
N (cm^{-3})	10^{15}	10^{15}	10^{15}	2×10^{19}	2×10^{19}	2×10^{19}	0	0	Channel doping
T_{fin} (nm)	9.6	8	6.4	9.6	8	6.4	5	10	Body/fin width
Ψ_M (eV)	4.47	4.47	4.47	5.25	5.02	4.82	4.15	4.15	Gate work function
V_{DD} (V)	0.9	0.81	0.73	0.9	0.81	0.73	1	1	Power supply voltage
$V_{T,lin}$ (mV)	272	282	298	306	306	300	*n/a*	*n/a*	Lin. threshold voltage
$V_{T,sat}$ (mV)	201	203	208	200	192	185	440	553	Sat. threshold voltage
I_{on} (μA/μm)	1432	1527	1734	1144	1225	1330	0.227	0.171	On current with $V_{GS} = V_{DS} = V_{DD}$
I_{off} (nA/μm)	6.7	9.7	13.3	11.3	21.3	36.4	0.19 fA/μm	0.134 fA/μm	Off current with $V_{GS} = 0$ & $V_{DS} = V_{DD}$
SS (mV/dec)	67.9	69.8	71.6	72.5	74.2	75.3	65.6 (avg.)	80.6 (avg.)	Subthreshold swing
DIBL (mV/V)	24.0	32.0	39.7	77.3	89.8	95.6	*n/a*	*n/a*	Drain-induced barrier lowering

Chapter 4, there is no design consensus for TFETs as of this writing, so we consider hypothetical "20/5" and "20/10" designs which emulate possible designs which are dimensionally similar to the FinFET generations being investigated. The exact design specifications are not important, since our variability results are eventually normalized against the baseline values in the lower half of Table 10. For the FinFETs, high field transport is captured with a calibrated hydrodynamic model and quantum corrections are modeled using the density gradient approximation (DGA). For TFETs, we use a calibrated nonlocal band-to-band tunneling (BTBT) model and the modified local density approximation for quantum corrections (MLDA). Reverse diode leakage is modeled using field-assisted Shockley-Read-Hall lifetimes.

To incorporate LER in our simulated devices, we generate random LER patterns based on a Gaussian model with root-mean-square roughness amplitude σ_{LER} = 1 nm and correlation length λ = 15 nm. These patterns are used to augment the body sidewalls in our devices resulting in random fluctuation of the body or fin width in each FET structure. LER along the gate line is not considered in this work and all LER profiles are assumed to be uncorrelated. In addition, RDF is incorporated in our structures by randomizing the position and number of ionized dopants according to a Poisson distribution as described in Chapter 3, and considering the long-range Coulomb potential from each discrete dopant in the Poisson equation with an appropriate screening length. Double-counting of screening effects from the long-range potential and the DGA is avoided in

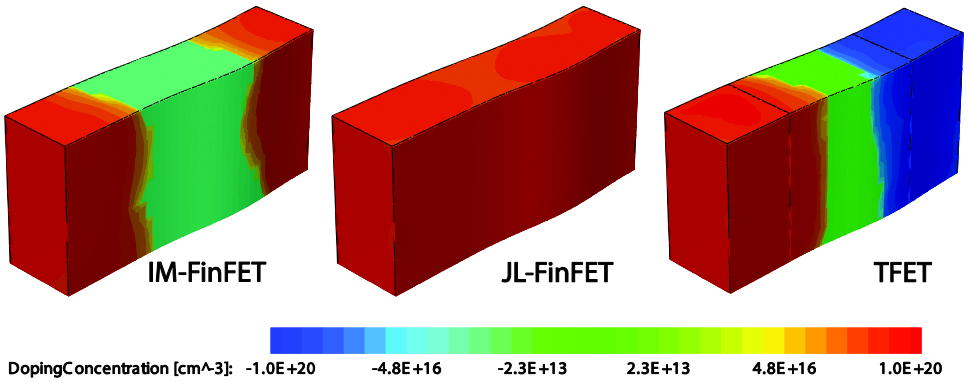


Fig. 38. Examples of simulated structures containing body/fin LER and RDF: (left) 32nm IM-FinFET, (center) 32nm JL-FinFET, and (right) 20/10 TFET. All devices are shown with a height of 20 nm.

Sentaurus Device since the atomistic doping profile is first transformed into a smoothened continuous profile before actual device simulation takes place, such that the DGA does not "see" any potential singularities near individual dopants and ultimately double-screen them. Example structures containing both LER and RDF are shown in Fig. 38. By simulating a statistically large number of devices (200 in our case) with LER and/or RDF, we obtain distributions in each of the performance figures $V_{T,lin}$, $V_{T,sat}$, I_{on}, SS, and DIBL for each transistor type. To make our results more concise, we will only use results for LER and RDF variability with σ_{LER} = 1 nm and a transistor height H = 10 nm, representing the worst cases in Chapter 2 and Chapter 3.

When the effects of multiple variability sources act independently, their combined impact on the standard deviation of some parameter, e.g., V_T, can be expressed by computing the root-squared sum (RSS) of the standard deviations from each variability source, as in

$$\sigma V_T = \sqrt{\sigma V_T^2(LER) + \sigma V_T^2(RDF) + \cdots}\,. \tag{4}$$

For each of the FET types shown in Fig. 38, we present in the following sections: a) RSS-predicted standard deviations of the six aforementioned performance figures when LER and RDF are each enabled *separately*, and b) standard deviations of the same performance figures as directly ob-

tained from simulations when LER and RDF are enabled *simultaneously*. By comparing the relative magnitudes of a) and b) within each class of transistor and calculating a percentage error associated with the RSS calculation (for V_T as an example):

$$\%\ Error = \left| \frac{\sigma V_T(RSS) - \sigma V_T(TCAD)}{\sigma V_T(TCAD)} \right| \times 100\% \tag{5}$$

we will determine whether the effects of LER and RDF may be viewed as independent or not.

5.3 IM-FinFET Joint Variability

Our first candidate is the IM-FinFET which is commonly believed to be robust against RDF because of its intrinsic channel[4], yet vulnerable to LER. In Table 11, we observe that RDF-induced variability is indeed mostly smaller than that due to LER, with the exception of σI_{on}. This is reasonable since LER primarily impacts short-channel effect (SCE) control in the fin which, at sub-32nm nodes, can be significant factor in device performance, especially at high drain biases. On the other hand, RDF modulates the source/drain resistance R_{sd} and effective channel length L_{eff} by an amount roughly given by twice the dopant screening length (about a few nanometers). For electrostatic-related metrics such as V_T and SS, this effect is marginal since the electrostatic gate control is not significantly compromised—hence the smaller variations in those parameters. For

Table 11. Inversion-Mode FinFET Variability from LER and RDF

Node	Source	$\sigma V_{T,lin}$	$\sigma V_{T,sat}$	σI_{on}	σI_{off}	σSS	σDIBL
32nm	LER	1.6%	5.3%	2.8%	76.3%	2.4%	24.8%
	RDF	1.3%	1.2%	3.3%	18.0%	0.6%	12.1%
21nm	LER	2.2%	7.5%	3.8%	104.1%	3.4%	29.0%
	RDF	1.0%	2.1%	5.5%	41.0%	0.6%	15.6%
15nm	LER	2.7%	10.8%	5.8%	227.2%	4.9%	38.5%
	RDF	3.6%	3.7%	15.3%	157.1%	1.2%	17.4%

[4] A residual background doping of ~10^{15} cm^{-3} in the silicon fin means there is still a remote chance of a single acceptor ion appearing in the channel with a probability of 1 in ~500 (1200) for 32nm (15nm) IM-FinFETs. Such an occurrence would most likely result in a "failed" device. Since our ensemble size is limited to 200 devices per technology, we did not see this occurring.

I_{on} which is largely transport-related, fluctuations in L_{eff} and R_{sd} become important, thus yielding higher σI_{on} from RDF compared to LER. Overall, however, the numbers in Table 11 suggest that variability due to LER and RDF (when treated separately) is manageable for IM-FinFETs.

Next, we compare the individual variability results with those when LER and RDF are simultaneously activated. The comparison is shown in Fig. 39(a) where the combined RSS-predicted variability from (4) is shown alongside the actual TCAD-simulated variability with simultaneous LER and RDF. The reader should ignore the "+4 nm overlap" data for now, as it will be used later. By assuming LER and RDF act independently, we have significantly underestimated the actual amount of variability that may be present, especially for $\sigma V_{T,sat}$, σI_{off}, and σDIBL. The percent errors calculated by (5) between the RSS and TCAD values in Fig. 39(a) are shown in Fig. 39(b) which illustrates the above fact. This predicted error exceeds 10% on average for all metrics, except SS. Based on the data shown we cannot conclude that the impacts of LER and RDF are statistically independent if an accurate estimation of variability from LER and RDF is needed.

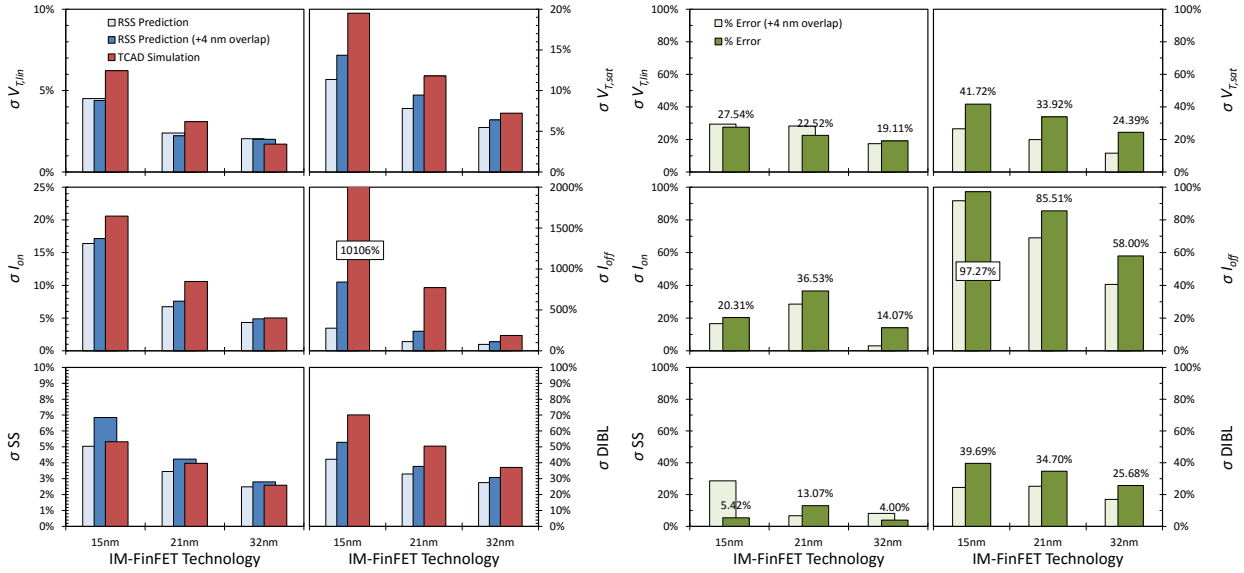


Fig. 39. (a) Comparison of expected IM-FinFET variability when LER and RDF are assumed uncorrelated versus direct simulations with LER and RDF present. (b) Percentage error incurred when assuming independence of LER and RDF compared to actual simulated values.

From a modeling standpoint, this is an unfortunate conclusion, since it implies that brute force device simulations may not be easily replaced by simple analytical descriptions, i.e., (4), for comprehensive variability assessment in IM-FinFETs.

There are several possible explanations for why (4) fails to capture the net performance variability from LER and RDF. The first centers on how fin LER and RDF affect the electrostatics in an IM-FinFET. When only LER exists along the channel, the local fin width fluctuates and affects the device's SCE control since T_{fin} now varies across the device, but L_{eff} remains unaltered. When only RDF exists, the junctions become less abrupt and encroach laterally into the channel (which reduces L_{eff}) and results in an unintentional overlap in most situations. When both LER and RDF exist simultaneously (as in Fig. 38), the loss in L_{eff} from RDF increases the device's sensitivity to fluctuations in T_{body}, meaning the device becomes more sensitive to LER when RDF is present. However, when LER and RDF are treated *separately* (as in the RSS prediction), the increased sensitivity is not accounted for, and we effectively neglect any interactions that LER and RDF may have—that which results in unexpectedly high variability from TCAD simulations compared to the RSS predictions. To highlight this phenomenon, we repeated the same comparisons as before, except with an additional 4 nm uniform overlap added to each junction in the LER-only simulations to crudely account for any possible LER-RDF interactions when comparing against the full TCAD results. The reader should note that this neglects nonuniformity in the overlap profiles which may occur from RDF. We can see from Fig. 39 that in most cases the RSS predictions with added overlap match the TCAD results more closely, although sizeable discrepancies still remain, especially for $\sigma V_{T,lin}$ and σI_{off}. Nevertheless, this data supports the first possible reason for the observed correlations between LER and RDF in IM-FinFETs.

Table 12. Suppressed LER-RDF Interactions in 15nm IM-FinFETs with L_g = 50 nm

Source	$\sigma V_{T,lin}$	$\sigma V_{T,sat}$	σI_{on}	σI_{off}
LER only	1.19%	3.03%	3.21%	24.00%
RDF only	0.36%	0.44%	4.35%	3.07%
LER+RDF (RSS)	1.24%	3.07%	5.40%	24.19%
LER+RDF (TCAD)	1.35%	3.23%	5.49%	25.46%
% Error	**7.69%**	**5.19%**	**1.52%**	**4.98%**

H_{fin} = 10 nm for simulated devices containing RDF.

Since we expect the unintentional overlap from RDF to be more consequential for short channel devices compared to long channel devices, we might anticipate the observed LER-RDF interactions to be weaker for IM-FinFETs with longer L_g. In Table 12, we see that for 15nm IM-FinFETs with an inflated L_g = 50 nm, the percent errors between RSS predictions and TCAD simulations are reduced to <10% for $\sigma V_{T,lin}$, $\sigma V_{T,sat}$, σI_{on}, and σI_{off}. Variations in SS and DIBL are not shown due to numerical artifacts during extraction, since all simulated values become approximately 60 mV/dec and 0 mV/V, respectively. From the data, we see that a reduction in L_{eff} by a few nanometers due to RDF has a much smaller impact when the nominal L_g is longer, so that any resulting LER–RDF interactions become suppressed.

The second possible explanation deals with the length scales over which LER and RDF manifest themselves in causing device variations. Since RDF only exists in the source and drain, its manifestation is "localized" to within nanometers of the source-channel and drain-channel junctions, and RDF has no effect along the channel midsection. On the other hand, fluctuations in T_{fin} due to LER appear everywhere along the fin, along with its impact on SCE control and overall device integrity, and so the significant impact of LER is spatially "distributed" over the entire device length (~ tens of nm). Both of these observations can be seen from Fig. 38. Since RDF only manifests itself at two locations (i.e., the junctions) in an IM-FinFET, it only interacts with LER at those same two locations. In other words, the convergence areas from LER and RDF are limited

in number and (relatively) small in scale. In the event that interactions occur between certain instances of LER and RDF, their combined effects will not average out the way they would if LER and RDF both affected device behavior over longer scales—this distinction will become evident in the following sections when we investigate the variability of JL-FinFETs and TFETs.

Given the above arguments, the most probable reason for the large discrepancies between RSS predictions and TCAD simulations (Fig. 39) is likely given by our first explanation: increased sensitivity to LER as a result of shortened L_{eff} from RDF overlap effects. Thus, for accurate IM-FinFET variability projections, LER and RDF effects must be modeled together, especially for short channel geometries.

5.4 JL-FinFET Joint Variability

Our next candidate is the JL-FinFET whose key difference from the IM-FinFET is the absence of *p-n* junctions along the channel, as shown in Fig. 38. With its uniformly high channel doping, small device dimensions, and depletion-mode operation, it is extremely vulnerable to LER and RDF. This fact is evident from Table 13, which reveals a much larger amount of performance variation for JL-FinFETs compared to IM-FinFETs (Table 11). Reasons for this have been thoroughly detailed in Chapter 3, and we will not repeat the same discussion here for brevity. What

Table 13. Junctionless FinFET Variability from LER and RDF

Node	Source	$\sigma V_{T,lin}$	$\sigma V_{T,sat}$	σI_{on}	σI_{off}	σSS	σDIBL
32nm	LER	40.9%	67.2%	19.6%	117005%	3.4%	26.8%
	RDF	44.5%	64.7%	20.8%	58265%	4.7%	51.8%
21nm	LER	36.0%	65.3%	20.5%	70713%	4.0%	24.3%
	RDF	36.9%	57.5%	18.9%	38570%	3.2%	33.1%
15nm	LER	31.6%	63.3%	22.2%	74151%	4.8%	25.9%
	RDF	26.1%	42.2%	22.3%	12239%	2.3%	22.4%

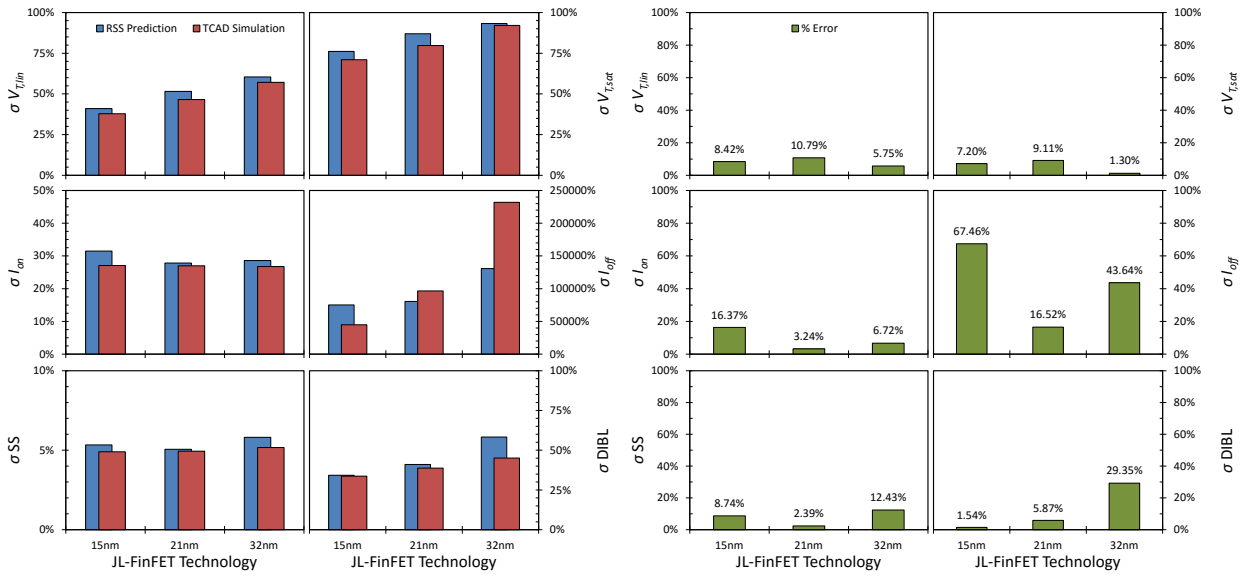


Fig. 40. (a) Comparison of expected JL-FinFET variability when LER and RDF are assumed uncorrelated versus direct simulations with LER and RDF present. (b) Percentage error incurred when assuming independence of LER and RDF compared to actual simulated values.

remains to be determined, instead, is whether LER and RDF show any obvious signs of interaction when simultaneously present.

In Fig. 40(a), we see that for JL-FinFETs the RSS-predicted values are generally much closer to the actual simulated values than they were for IM-FinFETs in Fig. 39(a). When percentage errors are compared again in Fig. 40(b) for JL-FinFETs, we see that most performance metrics incur errors of less than 10% on average except for σI_{off} which, although still large (due to its exponential nature), is smaller than for IM-FinFETs. Based on these results, it appears that the variability impacts from LER and RDF for JL-FinFETs can be considered independent with much greater confidence, unlike the case for IM-FinFETs. Unfortunately, an accurate estimation of σI_{off} is still problematic using the simple RSS formula—however, the difference between RSS-predicted and TCAD-simulated values are still roughly within a factor of 2× (130,000% *vs*. 230,000%) and well within an order of magnitude. By comparison, the largest difference in σI_{off} for 15nm IM-FinFETs is 36× (276% *vs*. 10,100%).

We might wonder why LER and RDF appear uncorrelated in JL-FinFETs, but not in IM-FinFETs. Previously for JL-FinFETs, we showed how fin LER can make the size and shape of the buried channel fluctuate to the extent that the channel may never fully open or close, depending on the actual line width roughness (LWR) along the channel direction. The same phenomenon occurs for JL-FinFETs when RDF is present—the size/shape of the buried channel undulates with the peaks and valleys in the electrostatic potential from the random placement of dopants. With this in mind, we recognize that both LER and RDF directly alter the shape/size of the buried channel in similar fashions and, when combined together, we can imagine how either mechanism can result in the accidental permanent opening or closing of a buried channel throughout the entire operating gate voltage range. Note that these possibilities are independent of SCE and may occur for any channel length, and that these mechanisms are inherent to their underlying technology.

In addition, the absence of *p*-*n* junctions along the channel direction and the "gated resistor" characteristic of JL-FinFETs mean the significant impacts of LER and RDF become distributed over the entire channel length, rather than localized at the source-channel junction as in the case of IM-FinFETs or TFETs, as we will see in the next section. As such, the sum of LER–RDF interactions tend to average out more so in JL devices and we can envision how their overall effects combine in an uncorrelated manner (Fig. 40), whereas they would not in IM devices (Fig. 39).

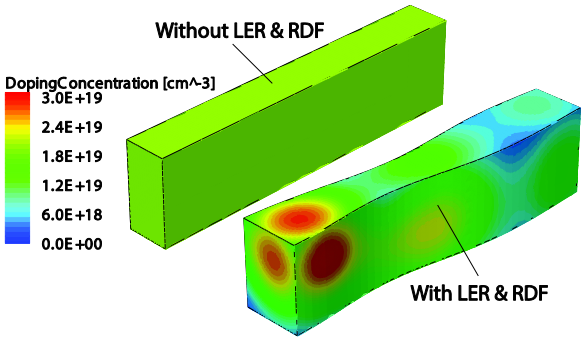


Fig. 41. Simulated resistors with and without LER & RDF. L = 40 nm, W = 5 nm, and H = 10 nm in the structures shown with σ_{LER} = 1 nm and nominal doping of 2×10^{19} cm^{-3}.

Table 14. Comparison of Resistor Current Variability from LER and RDF

Source	σI_R		
	10×5×10*	20×10×10*	40×20×10*
LER only	21.59%	10.17%	4.10%
RDF only	28.46%	18.81%	10.13%
LER+RDF (RSS)	35.72%	21.38%	10.93%
LER+RDF (TCAD)	34.85%	23.60%	10.58%
% Error	**2.50%**	**9.41%**	**3.34%**

*Resistor dimensions denoted by L×W×H (in nm). Sample size = 100. V_R = 1 V.

To better appreciate the above fact, we may extend the previous argument to the analysis of simple rectangular cuboid resistors with (ideally) uniform doping in the presence of LER (along one dimension) and RDF. In a resistor such as the one depicted in Fig. 41, the significant impacts of both LER and RDF are distributed equally across the length of the device since the doping strategy is homogeneous and the total resistance has no inherent bias to any particular region. In Table 14, we show that the impacts of LER and RDF on resistor current variation are essentially independent, with errors less than 10% for three different geometries under consideration. Given the similarity between simple resistors and JL-FinFETs in the "on" state, we can see how the distributed characteristic of LER and RDF for both cases results in seemingly uncorrelated behavior when the both LER and RDF are simultaneously present.

5.5 TFET Joint Variability

Our final candidate is the TFET whose fundamental operation is quite different from the IM- and JL-FinFET discussed previously. Comparing Table 15 with Table 11 and Table 13, we can see that performance variation from LER and RDF is generally worse in TFETs compared to

Table 15. TFET Variability from LER and RDF

Node	Source	$\sigma V_{T,sat}$	σI_{on}	σI_{off}	σSS
20/5	LER	7.8%	17.4%	423.4%	7.6%
	RDF	5.3%	18.1%	18005.7%	8.9%
20/10	LER	3.2%	12.2%	35.3%	4.2%
	RDF	2.3%	7.0%	6012.3%	4.2%

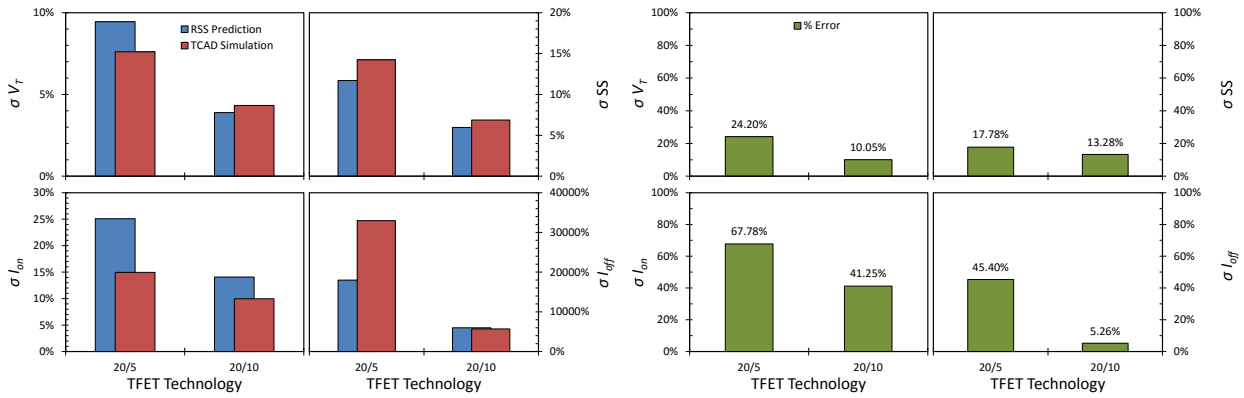


Fig. 42. (a) Comparison of expected TFET variability when LER and RDF are assumed uncorrelated versus direct simulations with LER and RDF present. (b) Percentage error incurred when assuming independence of LER and RDF compared to actual simulated values.

IM-FinFETs, but not nearly as bad as in JL-FinFETs. In Chapter 4, we explained why σI_{on} is strongly affected by LER and how RDF results in major degradation of σI_{off} resulting from loss of junction abruptness, causing the nominal I_{on}/I_{off} ratio to significantly worsen. We also showed preliminary evidence that TFET variability from LER and RDF did not combine independently when RSS-predicted values were compared with rigorous TCAD simulations, as depicted in Fig. 42. More than 10% error in σV_T, and over 40% error in σI_{on}, is obtained by assuming independence of LER and RDF in both 20/5 and 20/10 TFET designs; these are substantial indications that LER and RDF should not be considered independent for accurate estimations of these parameters.

To explain the apparent LER and RDF correlations in TFETs, we first note that nearly all of the transistor "action" in a TFET occurs at the source-channel junction where BTBT occurs. Because of this, the significant effects of LER and RDF are highly localized to the source-channel junction, and the only consequence of RDF is a loss in that junction's abruptness. As a result, the specific geometry and doping in other regions of the device (e.g., the main channel and drain sections) are secondary concerns at best; this is unsurprising given the predominant focus on source-channel junction engineering for TFET optimization in recent literature [56]–[60]. The signifi-

cance of this is that the exact geometry and profile of the source-channel junction becomes a critical factor in determining the overall device performance, such that any interactions between LER and RDF in TFETs are immediately apparent due to localization of variability effects.

The situation just described is in direct contrast to what we observed for junctionless devices. In the previous section, we noted that LER and RDF have a distributed effect over the entire channel length in JL-FinFETs where the transistor "action" (i.e. gate-modulated depletion) can be

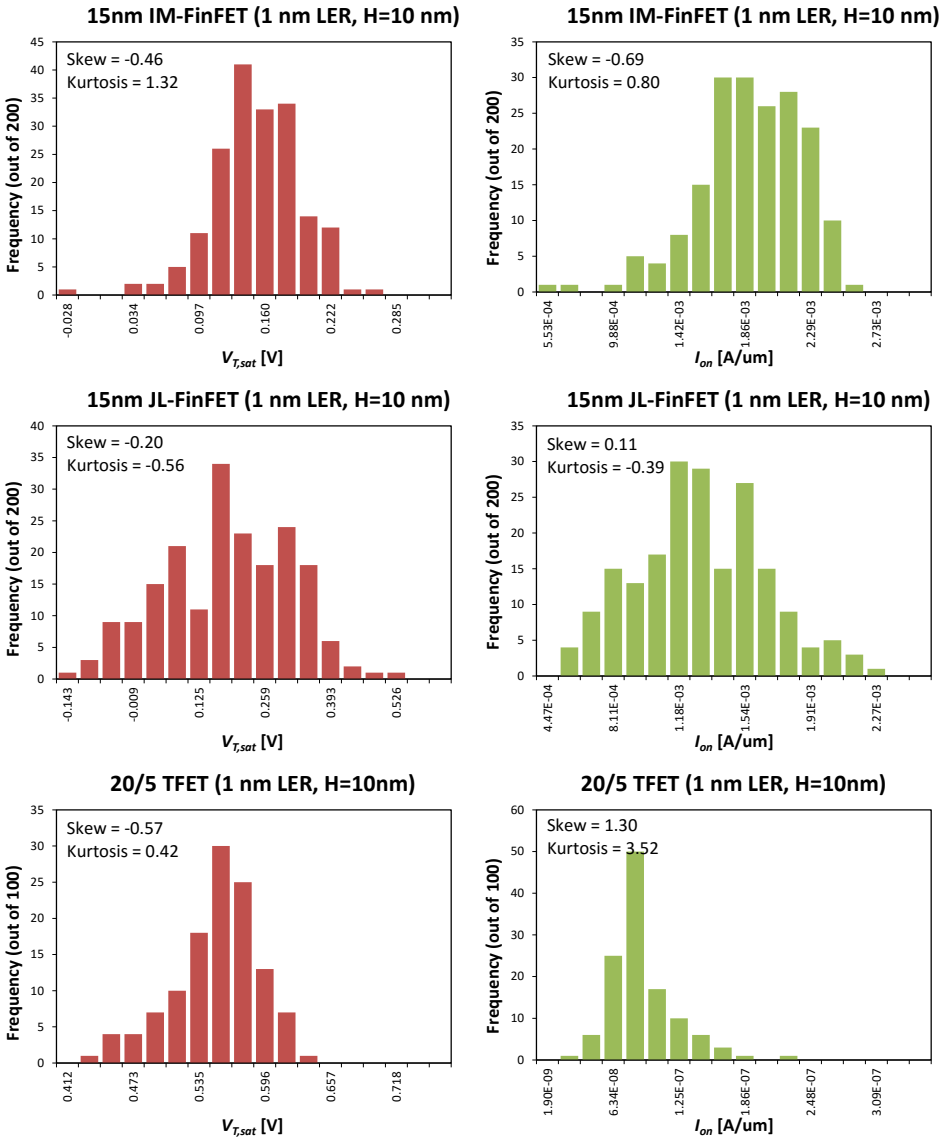


Fig. 43. Distributions of $V_{T,sat}$ and I_{on} for 15nm IM and JL-FinFETs and 20/5 TFETs with LER and RDF. The IM-FinFETs and TFETs have noticeable skew while JL-FinFETs appear normal symmetric.

impacted, and that LER and RDF cause performance variation in similar ways. Because of this distributed sensitivity, and the similar effects of LER and RDF, any nonlinear interactions between the two variability sources have the opportunity to average out over the channel length in junctionless devices, but not in TFETs where the effects are localized. As a final note, we observe in Fig. 43 that the distributions in $V_{T,sat}$ and I_{on} for TFETs (with LER and RDF) exhibit sizeable skew compared to IM-FinFETs and especially JL-FinFETs. The nonzero skew in TFET and IM-FinFET distributions are also suggestive of interactions between LER and RDF for those technologies.

5.6 Summary

The variability impacts of LER and RDF were investigated for IM-FinFETs, JL-FinFETs, and TFETs designed for 32, 21, and 15nm high-performance logic nodes. For JL-FinFETs, we have shown that LER- and RDF-induced variability combines in a statistically independent manner such that reasonably accurate estimations of device variability may be obtained from separate treatment of LER and RDF during simulations. By adding the individual variances in performance from LER and RDF (activated separately), and comparing the sum to the variance obtained by simultaneous treatment of LER and RDF during simulations, minimal error was found (< 10% in most cases) for σV_T and σI_{on} between the two approaches. However, an accurate estimation of σI_{off} still requires simultaneous treatment of LER and RDF, with an observed maximum error of 67% when independence is assumed. On the other hand, the same conclusions are not reached for IM-FinFETs and TFETs, wherein significant differences between the RSS-predicted and TCAD-simulated variations are observed. For IM-FinFETs, between 20–40% error is witnessed for σV_T and

σI_{on} and up to 97% (36×) error in σI_{off} is obtained when LER and RDF are assumed to be independent. For TFETs, between 40–60% error in σI_{on} is seen which represents the largest case of I_{on} estimation error among all technologies.

The opposing conclusions between junctionless and junction-based devices are qualitatively explained in terms of spatially distributed versus localized variability effects from LER and RDF in each FET technology. The lack of *p*-*n* junctions and depletion-mode nature of JL-FinFETs results in the significant effects of LER and RDF being distributed along the channel, allowing local LER–RDF interactions to average out across the device length. In the case of IM-FinFETs and TFETs, however, the reliance on a confined source-channel junction results in the significant effects of LER and RDF being localized near the junction vicinity, so that local LER-RDF interactions do not average out within a device. Consequently, the overall impacts from LER and RDF appear uncorrelated for junctionless devices and correlated for junction-based devices. With these findings, we conclude that a truly accurate projection of device variability for future IM-FinFET and TFET technologies requires a more comprehensive treatment of different variability sources, whereas JL-FET technology (despite the inherently larger variability) may be more predictable using independent treatments.

5.7 Appendix: Mean Parameter Shifts

For conciseness our data has, up to this point, been presented primarily in the form of computed standard deviations of $V_{T,lin}$, $V_{T,sat}$, I_{on}, I_{off}, SS, and DIBL and not the actual shapes of the statistical distributions. However, we have also independently examined the complete distributions of all six parameters for each permutation of FET technology and variability source in the same manner as Fig. 43. For brevity, the entire set of data is not shown here since the most significant findings have already been covered. As a final point, in Table 16 we present the calculated mean shifts in device performance (as a percentage of the baseline value given in Table 10) for each FET technology, which supplement the variability findings discussed in this transaction.

Table 16. Mean Parameter Shifts Relative to Baseline Values

LER	Node	$\mu V_{T,lin}$	$\mu V_{T,sat}$	μI_{on}	μI_{off}	μSS	μDIBL
IM-FinFET	15nm	-0.2%	-1.3%	-0.7%	79.2%	0.9%	6.4%
	21nm	-0.1%	-0.4%	-0.7%	31.4%	0.3%	2.2%
	32nm	0.1%	-0.2%	0.5%	14.0%	0.0%	2.1%
JL-FinFET	15nm	-5.1%	-11.8%	2.7%	12813.7%	0.8%	7.8%
	21nm	-2.0%	-7.1%	1.0%	15986.5%	-1.0%	7.9%
	32nm	2.9%	-0.7%	-0.7%	24742.1%	0.1%	11.6%
TFET	20/5	n/a	1.2%	-2.6%	136.6%	4.0%	n/a
	20/10	n/a	0.2%	1.9%	20.8%	1.2%	n/a

RDF	Node	$\mu V_{T,lin}$	$\mu V_{T,sat}$	μI_{on}	μI_{off}	μSS	μDIBL
IM-FinFET	15nm	-3.2%	-25.2%	7.5%	587.2%	5.9%	150.3%
	21nm	-0.2%	-11.9%	13.1%	185.8%	4.9%	82.7%
	32nm	-0.8%	-5.9%	8.7%	74.4%	2.7%	48.1%
JL-FinFET	15nm	-0.7%	-4.5%	-6.9%	2274.0%	0.6%	9.2%
	21nm	2.2%	-6.8%	-3.2%	8206.3%	0.4%	17.0%
	32nm	8.2%	-4.2%	-3.7%	15643.1%	2.8%	30.2%
TFET	20/5	n/a	15.3%	-48.9%	128330.6%	113.4%	n/a
	20/10	n/a	11.6%	-49.3%	51162.5%	89.3%	n/a

LER+RDF	Node	$\mu V_{T,lin}$	$\mu V_{T,sat}$	μI_{on}	μI_{off}	μSS	μDIBL
IM-FinFET	15nm	-6.2%	-33.2%	8.7%	3200.5%	3.8%	179.9%
	21nm	3.3%	-15.6%	17.9%	491.5%	3.7%	108.6%
	32nm	5.5%	-7.7%	24.4%	157.9%	2.6%	62.4%
JL-FinFET	15nm	-8.0%	-14.0%	-6.9%	15095.0%	1.0%	11.3%
	21nm	4.5%	-1.3%	-6.4%	25993.4%	0.4%	16.6%
	32nm	1.4%	-12.7%	-2.7%	77534.9%	2.2%	37.0%
TFET	20/5	n/a	16.5%	-52.4%	127770.8%	114.9%	n/a
	20/10	n/a	13.4%	-57.1%	48305.7%	91.1%	n/a

Chapter 6

Silicon vs. III-V Junctionless FET Variability[5]

6.1 Background

Although silicon-based technology remains the de facto standard in modern very large scale integrated (VLSI) systems—chiefly due to its low cost, excellent native oxide, and large wafer size—other material systems including Ge and many Group III-V semiconductors may offer theoretically superior performance in analog and digital applications. For example, in Fig. 44 and Fig. 45 we see that many III-Vs possess higher electron mobility and source injection velocity [73]

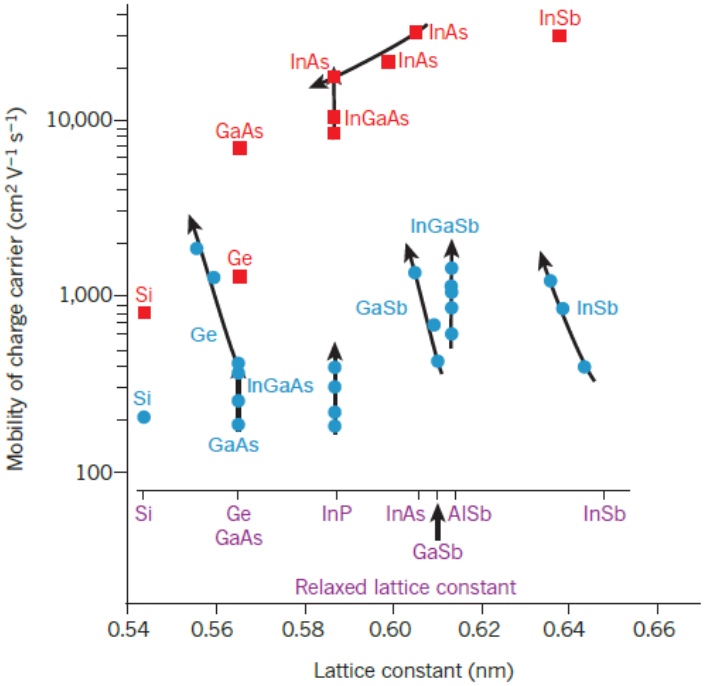


Fig. 44. Highest room temperature mobility of electrons (red) and holes (blue) versus semiconductor lattice constant in inversion layers and quantum wells. Data points which lie along a drawn arrow indicate different amounts of semiconductor biaxial strain and their respective strain-enhanced mobility. From [73].

[5] We sincerely thank Dr. Andrew Pan for his extensive contributions to this chapter. Most of the content in sections 6.2 through 6.5, and both appendices, were graciously prepared by Dr. Pan and reproduced in this chapter with his permission.

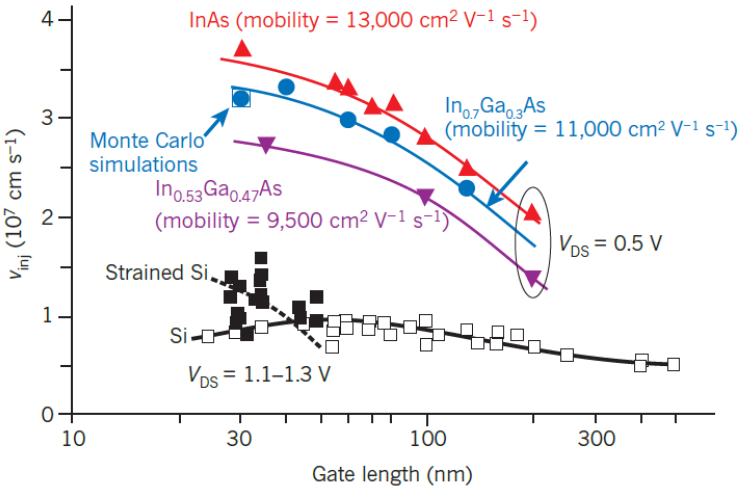


Fig. 45. Electron injection velocities of InGaAs and InAs HEMTs and Si MOSFETs as a function of gate length. The saturation of InGaAs channel mobility at shorter gate lengths indicates near-ballistic operation; this observation is supported by ballistic Monte Carlo simulations which lie coincident with the experimental data. From [73].

compared to Si, which make them promising candidates to replace *n*-Si FETs in nanoscale CMOS. Likewise, Ge and some Sb-based III-Vs are promising candidates to replace *p*-Si FETs thanks to greater hole mobility. Compared to Si, many III-Vs also have a significantly lower density of states (DOS) in the conduction band which—along with the greater drive current from faster carrier transport—can lead to lower gate capacitance and higher switching speeds in circuit applications. The lower conduction band DOS in III-Vs lead to other interesting properties which will be discussed later in the chapter as well.

Recently, a significant amount of theoretical and experimental research has been directed toward III-V FETs as leading candidates to replace Si in the nanoscale era. To this point, $In_{0.53}Ga_{0.47}As$ is arguably the most studied system due to its moderate band gap (0.74 eV), outstanding intrinsic electron mobility (>10^4 cm^2/Vs), low electron effective mass ($0.041m_o$), and because it can be grown lattice matched to InP. Many demonstrations of InGaAs MOSFETs have appeared over recent years, especially those utilizing nonplanar architectures such as FinFETs or GAA-FETs [74]–[83].

So far, however, even the best performing InGaAs FETs do not substantially outperform state-of-the-art Si FinFETs in terms of benchmarked I_{on}/I_{off} at V_{DD} = 0.5 V with a fixed $I_{off} \leq 100$ nA/μm. To our knowledge, the best performing Si and InGaAs FETs have been demonstrated from Intel®: their 14nm Si FinFETs (year 2014) [84] deliver $I_{on} \cong 0.45$ mA/μm at an I_{off} = 10 nA/μm, whereas their best InGaAs FinFETs (year 2011) [85] have shown $I_{on} \cong 0.375$ mA/μm at an I_{off} = 100 nA/μm. Currently, the major obstacle for short-channel InGaAs FETs is relatively high sub-threshold swing (often > 100 mV/dec at max drain bias), often attributed to a poor quality interface between the gate dielectric and III-V channel. A detailed investigation on the causes of poor interface quality is beyond the scope of this chapter, but it is clear that improving the SS of III-V FETs to levels <100 mV/dec will be paramount to the future III-V technology as a viable successor.

Despite the acceleration in research directed at III-V FET technology from the process development and design fronts, there is a lack of concrete understanding whether or not device variability will be better or worse for III-Vs compared to equivalent Si technology. Existing RDF studies on III-V devices typically focus on inversion-mode device operation [89] and/or lack meaningful comparisons against equivalently designed and operated Si devices [90]. Specifically, there is no clear understanding whether III-V based junctionless FETs are more or less vulnerable to RDF compared to equivalent Si-based designs.

This chapter then seeks to answer the question: will InGaAs or other III-V-based junctionless FETs be intrinsically more or less vulnerable to RDF than Si-based ones at the 15nm node when equivalently operated and designed? In order to answer this question, we must first consider the general implications of having a lower electron DOS in terms of local electrostatic response to potential variations. We elaborate upon this in the next section.

6.2 Effects of Degenerate Carrier Screening

Fig. 46 plots the conduction and valence band DOS for various semiconductors and shows that many III-Vs have a conduction band DOS that is roughly two orders of magnitude lower than in Si, meaning the onset of degeneracy occurs much sooner in III-Vs compared to Si. The relatively low conduction band DOS in many III-Vs leads to differences in electronic response when compared to higher DOS materials. When the DOS is low, the Fermi level must penetrate deeper into the band to populate the required electron density in the semiconductor. Now, if a local perturbation in the electrostatic potential dV appears in the semiconductor (either from an applied signal or outside source of fluctuation such as LER, RDF, etc.), the electron Fermi level changes by a proportional amount dE_F/q which leads to a subsequent change dn in the electron density. If the DOS is low, then dn will be relatively small because there are few states to populate/depopulate within an energy range of $\pm kT$ around the Fermi level[6] of a degenerate semiconductor. This means that the sensitivity of the carrier population, or effectively the electronic responsivity, to changes

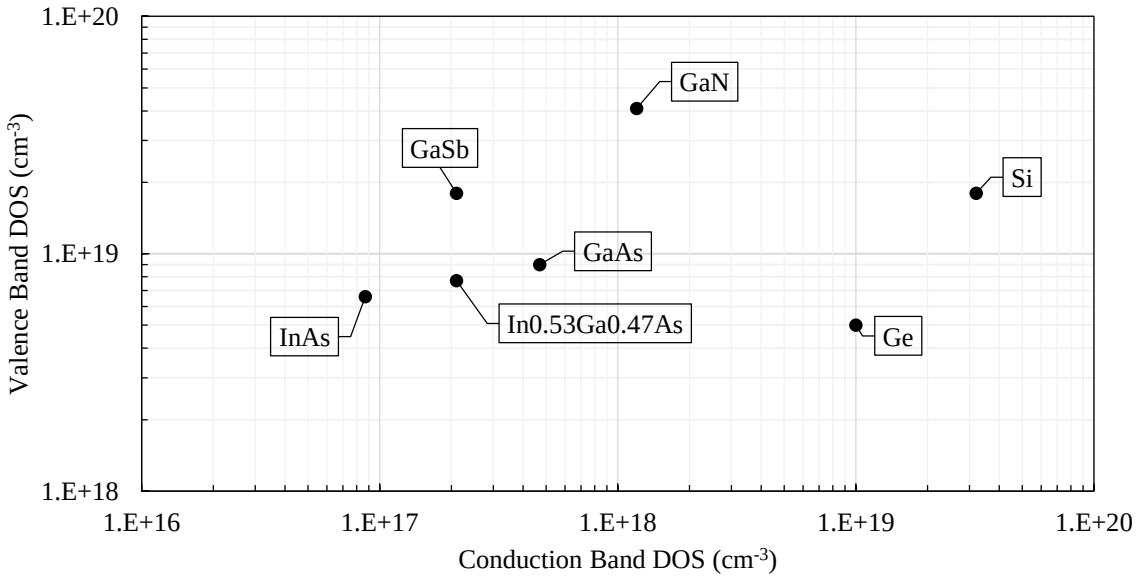


Fig. 46. Effective conduction band and valence band density of states in various semiconductors. Data taken from [91].

[6] Only electrons with an energy of $E_{Fn} \pm kT$ can respond effectively to perturbations because lower energy states are fully occupied/blocked according to the Pauli Exclusion Principle.

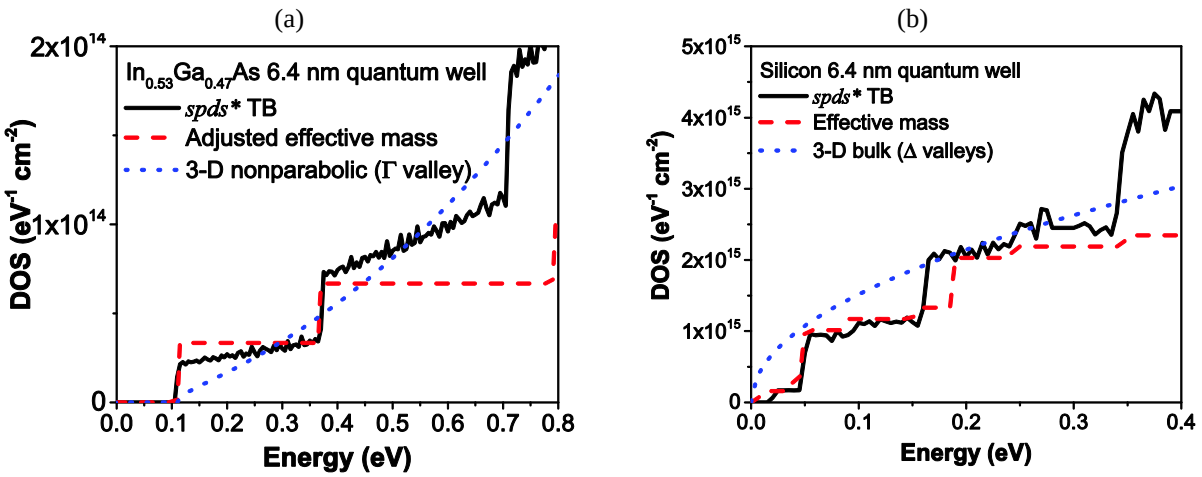


Fig. 47. Conduction band DOS in (a) $In_{0.53}Ga_{0.47}As$ and (b) Si quantum wells calculated using 2-D atomistic tight-binding (TB) and effective mass (EM) Hamiltonians, compared with the equivalent 3-D DOS normalized by the well thickness.

in applied voltage or sources of variation is innately weaker for degenerate semiconductors; in other words, dn/dE_F is lower for the degenerate case. Consequently, the effective screening length also becomes longer than the classical Debye length because of the weakened response.

The effects just described can be demonstrated in Si and InGaAs channels which will form the basis of study later in this chapter when we compare the variability of 15nm Si and InGaAs JL-FETs. In Fig. 47, we show the conduction band DOS for Si and InGaAs quantum wells (T = 6.4 nm) obtained from calibrated effective mass (EM) models and *spds** tight binding (TB) Hamiltonians [86], [87]. For reference, we also show the effective 2-D DOS which is defined as the 3-D bulk DOS divided by the channel thickness. In Si we used the bulk Δ valley masses, while for InGaAs we adjusted the EM to better fit the DOS from TB, with resulting values of $0.105m_o$ and $0.08m_o$ in the confined and unconfined directions respectively, where m_o is the free electron mass. Since perfect agreement between EM and TB calculations for all energies is impossible due to nonparabolic effects, we focus on fitting the EM DOS near the Fermi energies of the doping levels of interest in typical JL-FETs (N = 5×10^{18} to 4×10^{19} cm^{-3}). For InGaAs, this region lies between 0.2 to 0.4 eV.

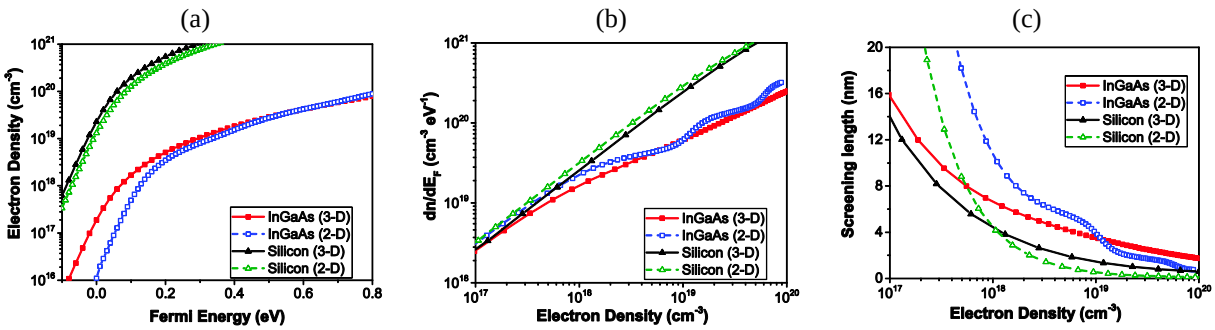


Fig. 48. Comparison of (a) n as function of Fermi energy E_F, (b) dn/dE_F versus n, and (c) screening length versus n in InGaAs and silicon. 2-D calculations are performed using the tight binding DOS. 2-D values of n and dn/dE_F are normalized to 3-D by dividing by the channel thickness T = 6.4 nm.

In Fig. 48(a), the electron density is shown as a function of Fermi energy for both Si and InGaAs in 3-D bulk and 2-D quantum well structures. The lower DOS in InGaAs causes E_F to rise quickly with n due to early onset of degeneracy. This can lead to significantly larger built-in voltages and will have implications for subthreshold variability in InGaAs JL-FETs, as we will see later. It also implies that the electron density responds more weakly to changes in local potential, *i.e.*, the degenerate dn/dE_F will be smaller than its nondegenerate counterpart [88]. We indeed observe in Fig. 48(b) that the InGaAs dn/dE_F becomes significantly smaller than in Si for $n > 10^{17}$ cm^{-3}, exactly the regime of interest for JL-FETs. This reduces the efficacy of free carrier screening, leading to longer screening lengths in InGaAs as seen in Fig. 48(c).

In the aforementioned plots we also compare the 3-D electron characteristics, computed using standard analytical formulas, with their 2-D counterparts obtained numerically using the TB DOS. As expected the trends, though quantitatively different, are qualitatively preserved across dimensions. The electrostatic effects of degeneracy affect the nominal device characteristics and also have a major impact on the resilience of different channel materials to device variability, as we will show later in the chapter.

6.3 Baseline Design & Performance of Si and InGaAs JL-FETs

Now that we have seen how degeneracy affects the electrostatics in Si and InGaAs channels differently, we return our attention to the scenario of 15nm Si and InGaAs JL-FETs which represent our case study of interest. The Si and InGaAs JL-FETs modeled in this chapter are based on the double gate JL-FinFET structure which we have examined previously in Chapter 3 and are dimensionally constrained according to the ITRS 15nm node [11] for high performance logic, thereby allowing a fair performance comparison between the two material systems assuming equal manufacturability in terms of lithography and process quality control. Table 17 lists the design values for the JL-FETs along with their respective nominal DC performance metrics, including linear and saturation threshold voltage ($V_{T,lin}$ and $V_{T,sat}$), on-state drive current (I_{on}), off-state leakage current (I_{off}), subthreshold. For both devices, the gate work function is chosen to obtain an I_{off} = 100 nA/μm.

Currently, there are no experimental demonstrations of JL-FETs (either Si or III-V based) which use a conventional symmetric double-gate structure at the nanoscale dimensions of interest,

Table 17. Nominal Parameters for Silicon and InGaAs JL-FETs

Quantity	NFET		PFET		Description
	Silicon	InGaAs	Silicon	InGaAs	
L_g (nm)	13	13	13	13	Physical gate length
EOT (nm)	0.64	0.64	0.64	0.64	Equivalent oxide thickness
N (cm^{-3})	2×10^{19}	2×10^{19}	2×10^{19}	2×10^{19}	Body doping
T (nm)	6.4	6.4	6.4	6.4	Body thickness
Ψ_M (eV)	4.730	5.130	4.505	4.630	Gate work function
V_{DD} (V)	0.73	0.73	0.73	0.73	Supply voltage
$V_{T,lin}$ (V)	0.238	0.264	-0.269	-0.262	Lin. threshold voltage (max g_m method with $\lvert V_{DS}\rvert$ = 50 mV)
$V_{T,sat}$ (V)	0.145	0.130	-0.143	-0.133	Sat. threshold voltage (constant $I = W/L_g\times10^{-7}$ A with $\lvert V_{DS}\rvert = V_{DD}$)
I_{on} (mA/μm)	1.60	2.82	3.70	4.04	On-state drive current with $\lvert V_{GS}\rvert = \lvert V_{DS}\rvert = V_{DD}$
I_{off} (nA/μm)	94.5	138	88.3	127.9	Off-state leakage current with $\lvert V_{GS}\rvert$ = 0 & $\lvert V_{DS}\rvert = V_{DD}$
SS (mV/dec)	70.2	69.7	69.7	70.6	Subthreshold swing
DIBL (mV/V)	67.0	80.4	64.1	71.7	Drain-induced barrier lowering

so we rely on quantum transport simulations based on the non-equilibrium Green's function (NEGF) formalism [92] to obtain the most physically accurate device model as our starting point. For brevity, we forego an in-depth review behind the concepts of NEGF since it is beyond the scope of this chapter and is not critical to our understanding of the variability trends that will be revealed in the next section. Rather, in this section we cover just enough details about the nominal JL-FET simulations performed using in-house NEGF code (courtesy of Dr. Andrew Pan) to establish a performance baseline which is as realistic as possible for the 15nm node. Once a set of baseline *I-V* curves are obtained for the Si and InGaAs JL-FETs from NEGF, we use them to calibrate standard drift diffusion models within TCAD for the variability analysis to follow.

The self-consistent NEGF simulations used for our JL-FETs include calibrated band structure as well as impurity, phonon, and surface roughness (SR) scattering models using techniques which are explained in Appendix I of this chapter. In Fig. 49, we show the nominal transfer curves for 2×10^{19} cm^{-3} doped *n*-type Si and InGaAs JL-FETs computed under ballistic and scattering conditions, and in Table 18 we list the corresponding values of I_{on} obtained under the different scattering conditions. The SR parameters are assumed to be Δ = 0.4 nm and 1.76 nm for Si and

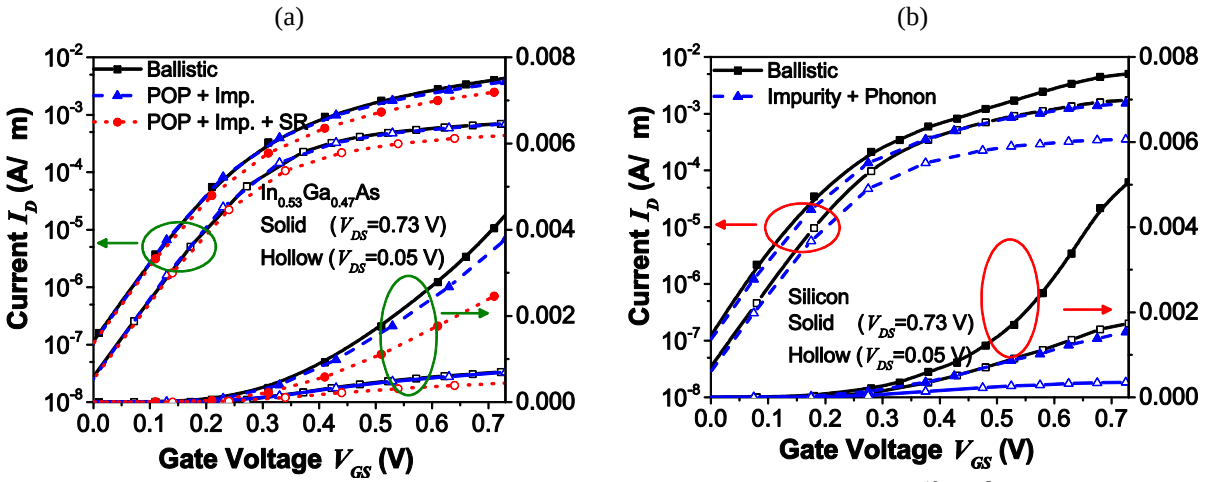


Fig. 49. Device characteristics for (a) InGaAs and (b) silicon n-type JL-FETs with 2×10^{19} cm^{-3} channel doping from ballistic (solid lines) and scattering (dashed lines) NEGF simulations. For InGaAs, the curve including SR scattering is calculating assuming a roughness amplitude Δ = 1.76 nm.

Table 18. Simulated I_{on} (in mA/μm) for 2×10^{19} cm^{-3} JL-FETs with Different Scattering Models.

Scattering Models	Silicon	InGaAs
None (Ballistic)	5.06	4.34
Impurity	1.58	3.77
Impurity + Phonon	1.53	3.70
Impurity + Phonon + SR (0.4 nm)	1.51	3.45
Impurity + Phonon + SR (1.76 nm)	-	2.47

InGaAs, respectively with $\lambda = 2$ nm in both cases (see Appendix I). Our results clearly show that impurity scattering has a major impact on the performance of Si devices, leading to over a 60% decrease in I_{on}. This is on par with the reduction observed in Si JL-FET Monte Carlo simulations when scattering is included [93]. By contrast, the InGaAs device is less affected by impurity scattering and I_{on} is only reduced about 10%; this is due to its much higher material mobility, as shown in Table 22. Previous ballistic studies of Si and III-V JL-FETs have concluded that the latter have lower currents due to low DOS limitations [94]; however, we observe that the greater mobility degradation in Si reverses this trend and leads to a larger I_{on} in equivalently doped InGaAs devices. For both devices, phonon scattering plays a negligible role as indicated in Table 18. SR scattering is also negligible for Si at $\Delta = 0.4$ nm, but is non negligible for InGaAs at $\Delta = 1.76$ nm[7]. Therefore, accurate performance assessments of JL-FETs must consider scattering effects.

The high Fermi energy in InGaAs devices arising from the small DOS and reduced quantum capacitance also alters their electrostatic behavior compared to Si. We illustrate this effect in Fig. 50, comparing the off- and on-state band diagrams and energy-resolved current along the center of the channel where the electron density is highest. When the device is off, we observe that the strong Fermi degeneracy in the source and drain of InGaAs devices leads to substantially higher

[7] We should keep in mind, however, that the values of Δ assumed here are taken from model fits against recent experimentally measured Si and InGaAs ultrathin channel mobility data and may not accurately represent the physical roughness in real systems, nor does it necessarily represent the expected quality for a hypothetically mature InGaAs technology. This issue is discussed further in Appendix I.

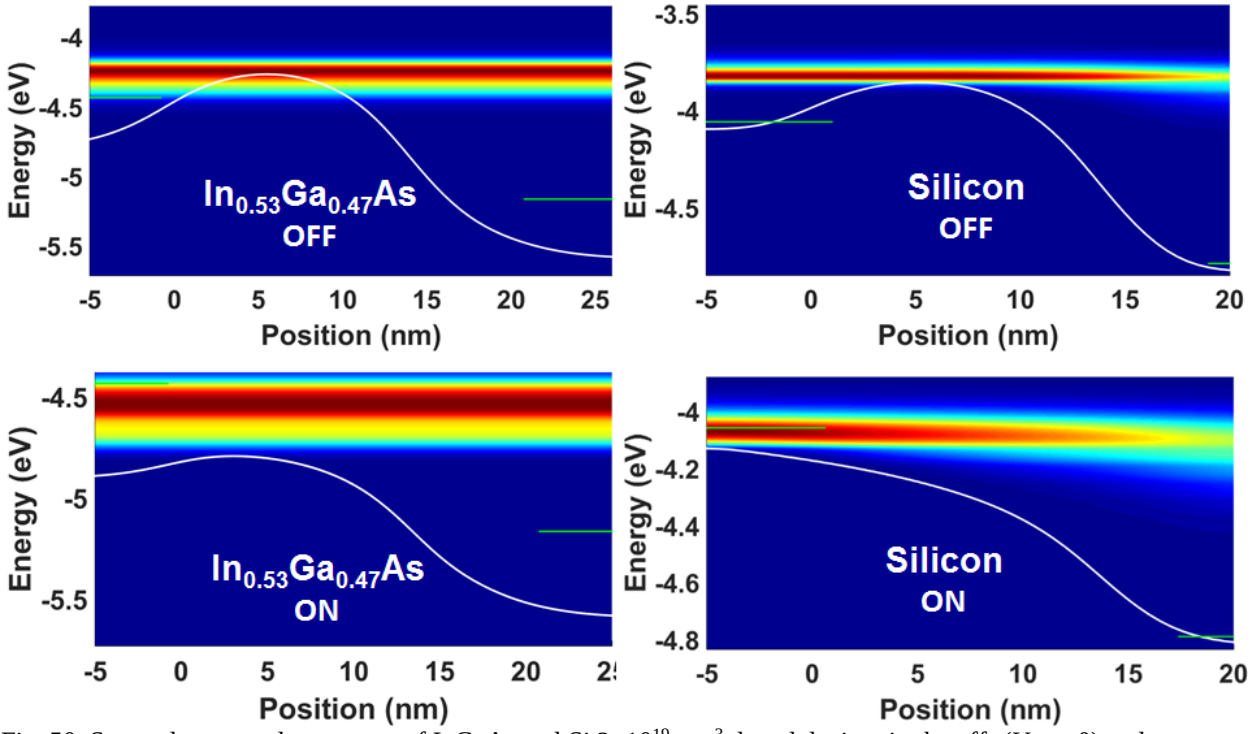


Fig. 50. Spectral current along center of InGaAs and Si 2×10^{19} cm^{-3} doped devices in the off- (V_{GS} = 0) and on-states (V_{GS} = 0.73 V). The green lines indicate the position of the source Fermi energy and white lines mark the first subband edge. Note the different energy scales for InGaAs and silicon.

channel electrostatic barriers compared to Si. As a result, in the on state, we find that the channel potential barrier has "collapsed" for Si but remains in the InGaAs device. Note that the majority of the current flows over the barrier even in the off-state, indicating that source-drain tunneling is not yet a major concern for InGaAs JL-FETs at this doping. As devices continue to scale down or doping increases, however, tunneling will become more important.

We must point out that while the *n*-type JL-FET simulations include scattering, the *p*-type JL-FET simulations using 6-band k-p Hamiltonians are ballistic because the more complex valence band models make scattering simulations impossible with the computational resources available to us. Later, we will show that the calibration using ballistic simulations does not affect our overall findings for the variability comparison between Si and InGaAs JL-FETs in the next section.

Table 19. Calibrated TCAD Parameters

Quantity	*n*-Si (*scattering*)	*n*-InGaAs (*scattering*)	*p*-Si (*ballistic*)	*p*-InGaAs (*ballistic*)	Description
μ (cm^2/Vs)	100	180	250	220	Low-field mobility
v_{sat} (cm/s)	2×10^7	8×10^7	5×10^7	3.5×10^7	Saturation velocity
β	1	2	1	2	Critical field exp.
m/m_o	*n/a*	0.42	*n/a*	*n/a*	Effective DOS mass
ΔE (eV)	*n/a*	0	*n/a*	*n/a*	Valley energy shift
d	*n/a*	1	*n/a*	*n/a*	Valley degeneracy
α	*n/a*	1.224	*n/a*	*n/a*	Nonparabolicity

With the nominal JL-FET transfer curves in Fig. 49 now established, we take a semiclassical drift-diffusion model (DD) in Sentaurus TCAD [10] and calibrate the transport parameters listed in Table 19 against the NEGF data. For InGaAs, Fermi-Dirac statistics and conduction band nonparabolicity (only for *n*-InGaAs) are included when calculating the electron density, whereas Boltzmann statistics and parabolic bands are assumed for Si. No quantum correction models are enabled in the TCAD simulations since none of the available models have calibrated parameters for materials other than Si, and we found through independent trial and error that tuning said parameters for InGaAs could not yield a better fit against the NEGF data compared to using no quantum model at all. For junctionless devices, we expect that neglecting quantization will not be as problematic since the channel naturally forms at the midsection of the body as opposed to near the oxide-semiconductor interface as in traditional inversion-mode devices. A more important concern is that in thin-body devices, the reduced dimensionality changes the DOS from its bulk (3-D) form, which in turn affects carrier screening and other electrostatic properties. Semiclassical TCAD is based on 3-D carrier statistics and cannot capture this effect quantitatively. However, the qualitative differences between materials which lie at the heart of this study, in particular the reduced DOS, screening, and dn/dE_F of InGaAs compared to Si, carry over from 3-D to 2-D. Therefore we believe the trends uncovered in our TCAD simulations will be representative of those found in lower dimensional structures as well.

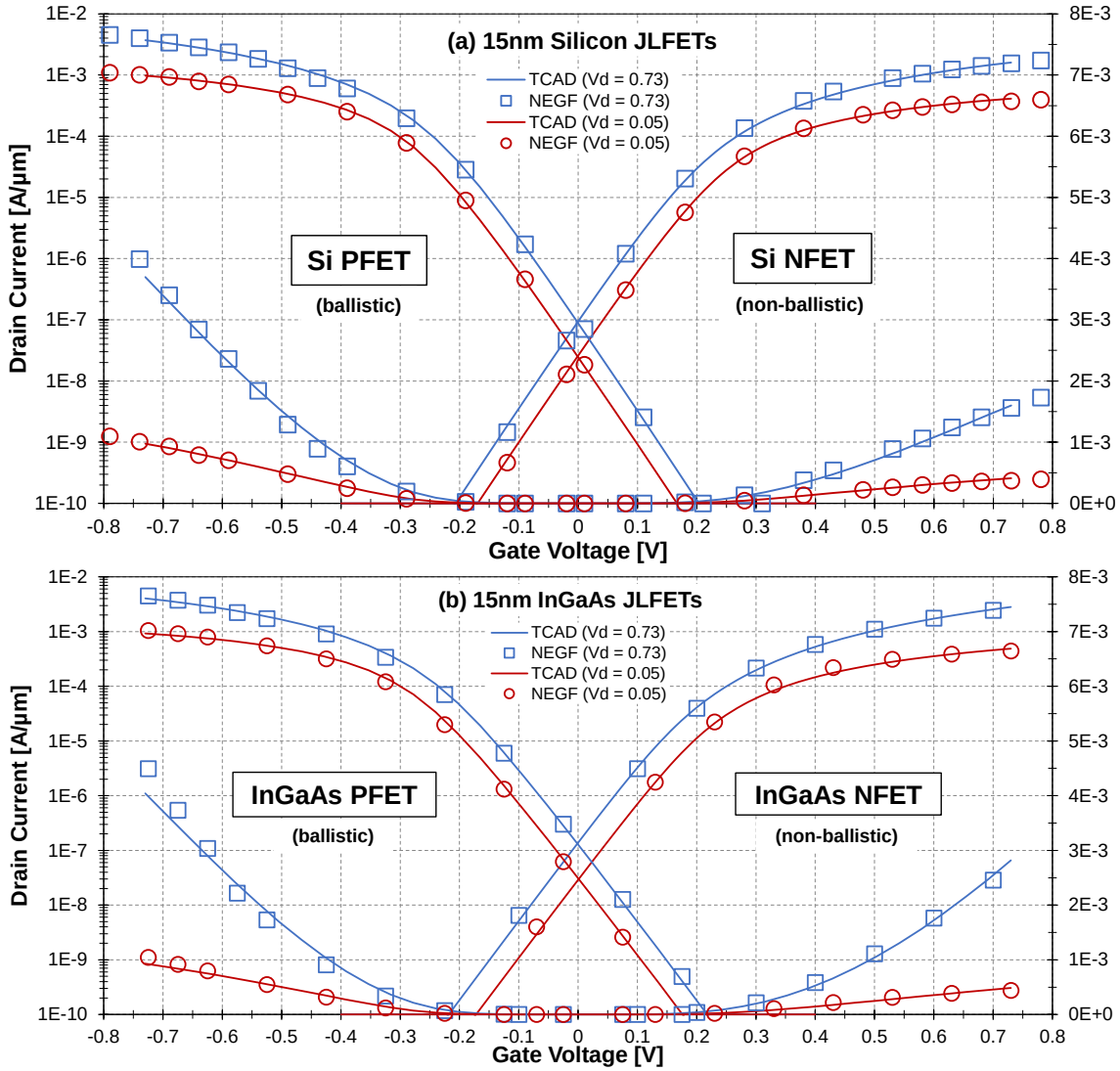


Fig. 51. Nominal I_D-V_G curves for 15nm (a) Si and (b) InGaAs JL-FETs showing TCAD calibrations performed against NEGF simulations. The upper curves in each panel correspond to the log scale on the left while the lower curves correspond to the linear scale on the right.

Tunneling processes (of any kind) are also not considered in this work since our NEGF simulations indicate that direct source-to-drain tunneling is negligible at 2×10^{19} cm^{-3} doping. The nominal transfer curves for our Si and InGaAs JL-FETs are displayed in Fig. 51 along with the calibrated fits against NEGF simulations.

6.4 Modeling RDF: Approach and Limitations

To capture the effect of RDF in our uniformly 2×10^{19} cm^{-3} doped JL-FETs, random local doping profiles are generated according to the Sano method [25] wherein discrete dopants are assigned to random locations within the device following a Poisson distribution. The effect of RDF is modeled via the long-range Coulomb potential established by individual dopants—each possessing a cutoff length $1/k_c = 1/2N(x,y,z)^{1/3}$ where N is the local impurity density located at position (x,y,z). This results in effectively non-uniform, random doping profiles within different JL-FET instances as depicted in Fig. 52. Device simulations are then performed on an ensemble of JL-FETs with random doping profiles to obtain values for $\sigma V_{T,lin}$, $\sigma V_{T,sat}$, σI_{on}, σI_{off}, σSS, and σDIBL resulting from RDF variability.

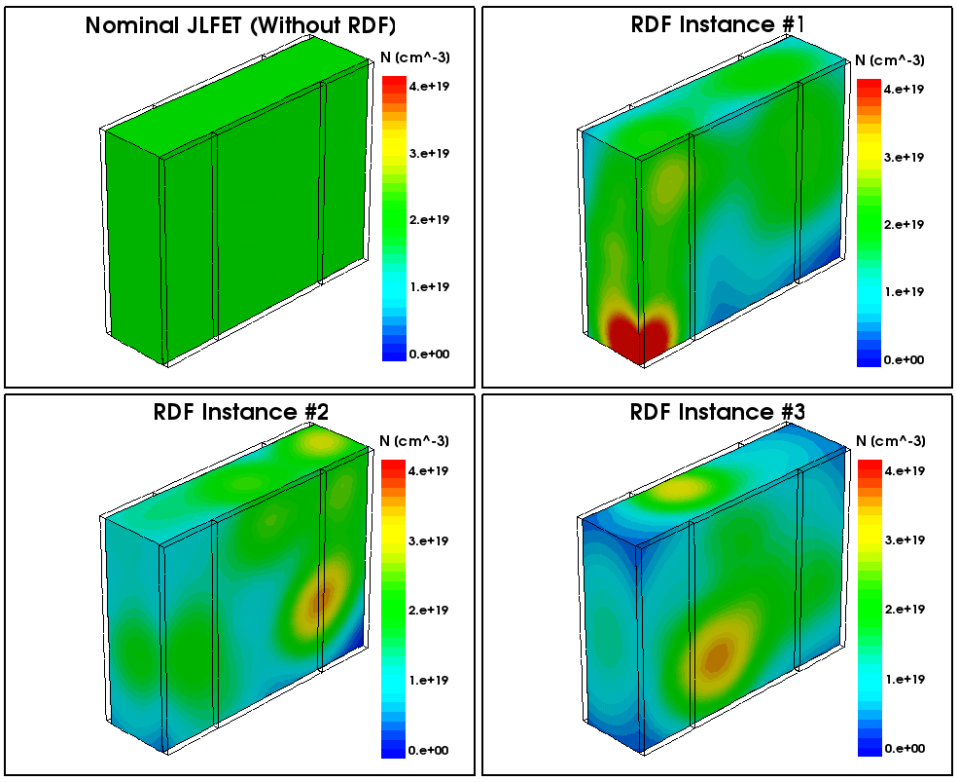


Fig. 52. Examples of JL-FETs exhibiting RDF generated from the Sano method. For reference, the nominal structure (without RDF) is shown in the upper left panel, having a uniform doping concentration of 2×10^{19} cm^{-3}.

There is ongoing debate over how to best treat discrete dopant effects in TCAD simulations since different approaches entail their own merits and limitations [25], [95], [96]. Perhaps the most rigorous and physically correct approach to treat RDF is to rely on true quantum mechanical simulations (e.g., NEGF) with discrete dopants placed on a finite mesh (the "atomistic" approach). However, the computational burden of simulating a statistically robust number (hundreds or more) of random devices using this method is often prohibitive.

Atomistic DD simulations have been known to produce incorrect carrier densities due to artificial charge trapping and other physical inconsistencies such as mesh size dependencies [95]. An efficient, yet still atomistic, method is to use DD simulations with the density gradient approximation (DGA) to avoid mesh size dependencies. This combination can still result in quantitative discrepancies compared to macroscopic theory, however, unless empirical corrections are made to the material parameters [96]. As mentioned before, any use of quantum correction models—including the DGA—did not result in a suitable fit for our nominal InGaAs characteristics from NEGF.

Sano's method avoids mesh size dependencies and does not require the use of any quantum correction models, however the proper choice of cut-off parameter k_c is ambiguous. In principle, k_c can be fitted to produce results which match macroscopic theory, but this may or may not match well with a truly self-consistent quantum mechanical solution in the microscopic domain. Unfortunately, there is no clear solution to this problem so the quantitative variability numbers in this work should be taken with the same degree of caution as in other semiclassical studies.

Since we neglect tunneling in our TCAD simulations, we neglect the possibility that local doping fluctuations may introduce additional direct tunneling in the off state, particularly in InGaAs if a large portion of the channel exceeds 4×10^{19} cm^{-3} doping. In theory, this could result

in an understimation of the true subthreshold variability for InGaAs; however, we anticipate that this effect will be minor because it is highly unlikely that a sufficiently large portion of the channel will reach such high doping to introduce substantial tunneling current. In the RDF structures of Fig. 52, for example, most of the channel hovers around 2×10^{19} cm^{-3} with only sparse pockets of higher doping. Nevertheless, we cannot discount said possibility as another limitation in our study.

Lastly, we have also decided not to incorporate any explicit doping dependence in the low field mobility values for Si and InGaAs JL-FETs since: 1) the process of calibrating against NEGF may compromise any physical meaning behind conventional doping dependent mobility data, and 2) it is more difficult to calibrate a doping dependent mobility model than it is to simply tune the low field mobility. The calibrated *n*-Si mobility of 100 cm^2/Vs is in fact very close to experimentally measured bulk values at $N = 2\times10^{19}$ cm^{-3} [97], however the calibrated μ and v_{sat} values in Table 19 for *p*-Si and InGaAs are understandably quite different from accepted bulk values. The main shortcoming of this approach is that a constant low field mobility which is independent of local doping fluctuations is assumed, which may underestimate device variability.

6.5 RDF in Doped Semiconductor Slabs

Before we present the variability results for our Si and InGaAs JL-FETs, it is useful to first examine how RDF causes local potential and carrier density fluctuations in uniformly doped semiconductor slabs. Although simple, this exercise will help us compare the impact of RDF in degenerate and nondegenerate semiconductors and will prove useful in explaining the JL-FET variability findings in Section 6.6.

In Fig. 53, the potential and electron density along the center cutline of a 100×100×100 nm^3 cuboid resistor (made of *n*-Si or *n*-InGaAs) with a given fixed random doping profile is shown.

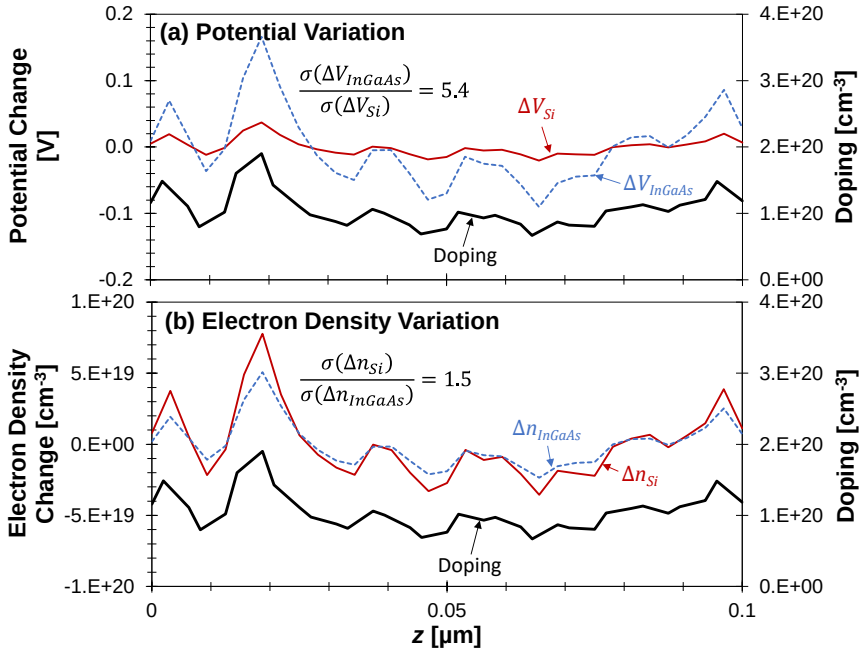


Fig. 53. Comparison of spatial fluctuations along a *z*-cutline in (a) electrostatic potential and (b) electron density in 100×100×100 nm^3 *n*-Si and *n*-InGaAs resistor slabs resulting from RDF. The nominal doping concentration (without RDF) for both slabs is 10^{20} cm^{-3}. Both slabs have exactly the same number and spatial arrangement of dopants.

The nominal doping in the slab is 10^{20} cm^{-3} and the slab is kept in equilibrium. We see that the potential fluctuates ~5.4× more in the InGaAs slab compared to the Si one, while the electron density fluctuates ~1.5× less in InGaAs compared to Si. This is a direct result of the longer screening length in degenerate InGaAs compared to nondegenerate Si as witnessed in Fig. 48(c).

In addition, we simulated ensembles of smaller 20×20×20 nm^3 resistors with different doping and extracted the integrated root-mean-square potential and electron density variations, with the results shown in Fig. 54(a)–(b). Again, the degenerate InGaAs slabs exhibit stronger potential fluctuations but also weaker electron density fluctuations when compared to Si. The ratio differences at 2×10^{19} cm^{-3} doping are roughly 5× and 1.5× for potential and carrier density, respectively, which are consistent with those of the single large slab in Fig. 53.

The fluctuations in slab current with a 10 mV bias applied are shown in Fig. 54(c). We see that the InGaAs slabs have roughly 11× higher current variation compared to Si, which is reconciled by the low field mobility ratio of 16000:1400 = 11.3:1 between InGaAs and Si when doping-

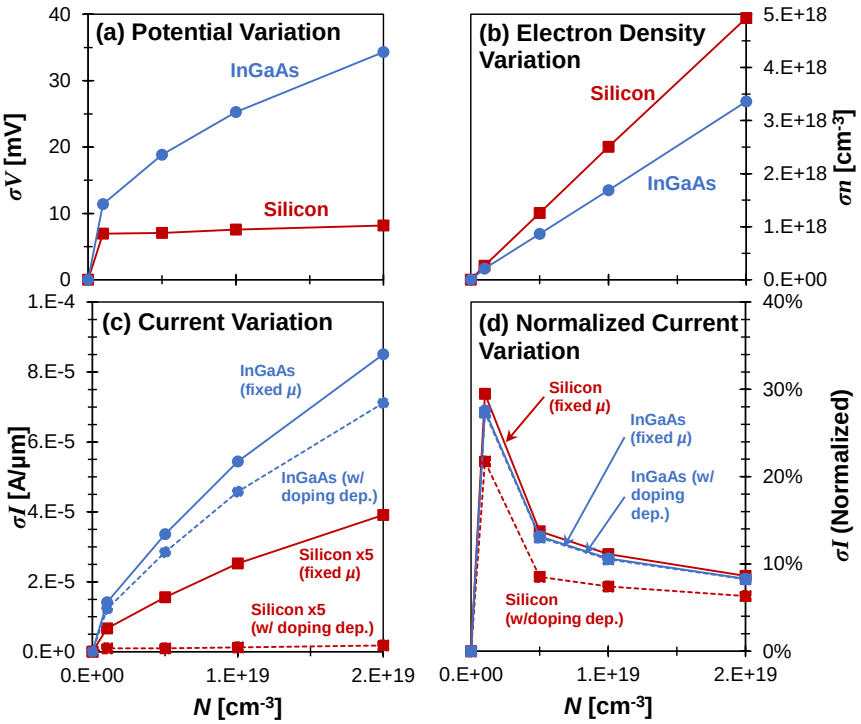


Fig. 54. Average fluctuations of (a) potential, (b) electron density, (c) current, and (d) normalized current in ensembles of 20 × 20 × 20 nm^3 Si and InGaAs slabs with RDF for different nominal doping concentration N. The ensemble size is 100 slabs for each combination of material and N. The applied voltage is 10 mV in (c) and (d). In (c), the Si curves are scaled by 5× for visual clarity. In (d), the current fluctuations are normalized to the ideal current values when RDF is absent from the slab.

dependent mobility is ignored. Interestingly, the relative variations (when normalized to their baseline values) in net slab conductance become nearly identical between InGaAs and Si. This shows that the two consequences of degenerate screening—higher potential fluctuations and lower carrier density fluctuations—effectively balance one another as far as carrier transport is concerned when the impurity and electron (average) concentrations are equal.

When doping dependent mobility is enabled using bulk parameter values, we see from Fig. 54(c)–(d) that the Si slabs exhibit less current variation compared to InGaAs. Since mobility μ varies inversely with doping, impurity scattering effectively counterbalances any local fluctuations in carrier density n from RDF (i.e., n and μ vary in opposite directions which acts to "stabilize" the product $n\mu$), thereby resulting in smaller variations in slab resistivity. Because impurity scattering effects are weak in InGaAs, the smearing of local resistivity is more evident in Si than in

InGaAs. Coupled with the lower potential fluctuations, and hence, lower built-in field variations in Si compared to InGaAs, it then makes sense that net fluctuations in slab current will be lower for Si than InGaAs.

6.6 RDF in Silicon and InGaAs JL-FETs

To compare the RDF-induced variability in Si and InGaAs JL-FETs, we examined a set of 200 devices for each combination of doping {*n*, *p*} and channel material {Si, InGaAs}, and compute the standard deviations in $V_{T,lin}$, $V_{T,sat}$, I_{on}, I_{off}, SS, and DIBL. The raw and normalized variability results are shown in Fig. 55 and Fig. 56, respectively. For the normalized results in Fig. 56,

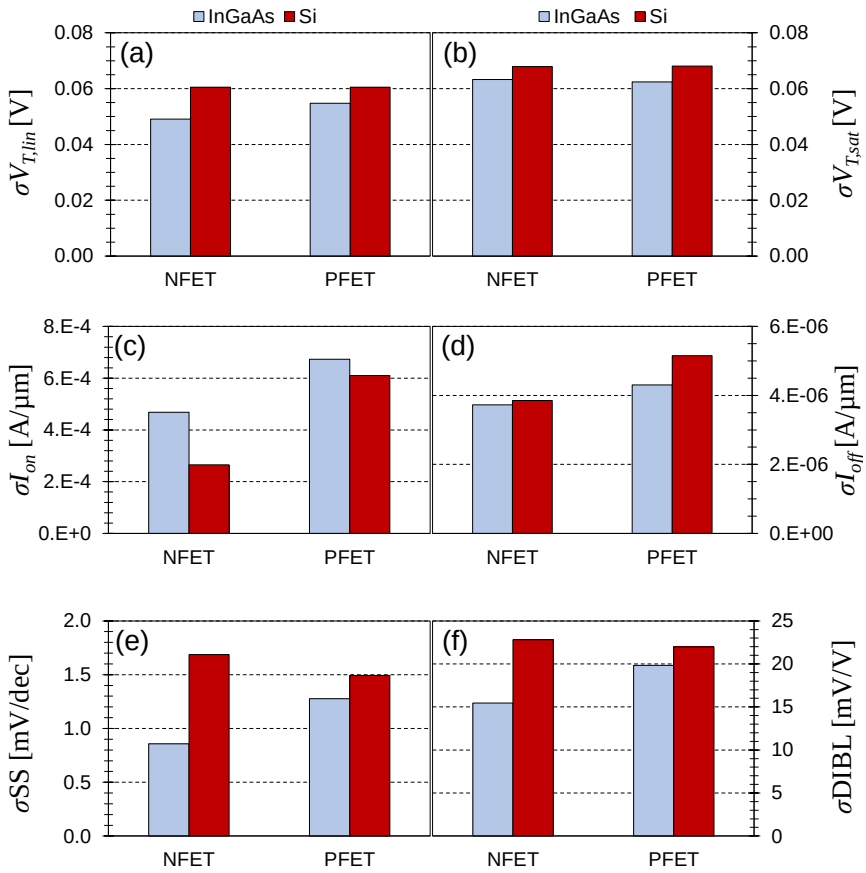


Fig. 55. Comparison of raw *n*-type and *p*-type InGaAs and Si JL-FET variability due to RDF for the metrics $V_{T,lin}$, $V_{T,sat}$, I_{on}, I_{off}, SS, and DIBL.

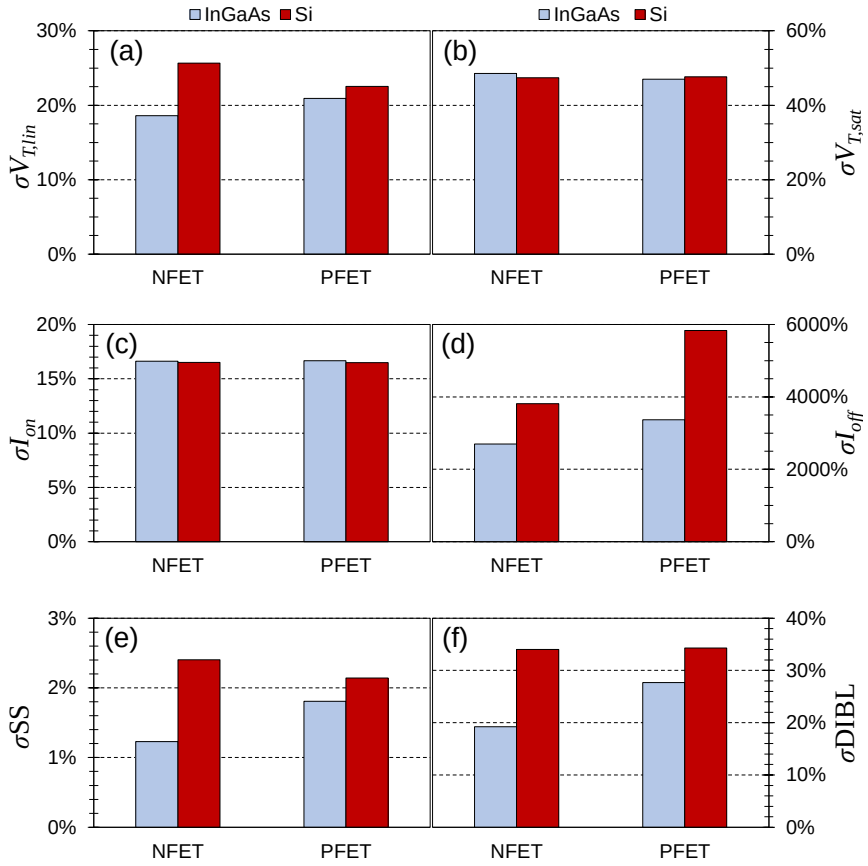


Fig. 56. Comparison of normalized *n*-type and *p*-type InGaAs and Si JL-FET variability due to RDF for the metrics $V_{T,lin}$, $V_{T,sat}$, I_{on}, I_{off}, SS, and DIBL. The standard deviations are normalized to the baseline values given in Table 17.

the standard deviation in each performance metric is expressed as a percentage of its nominal value from Table 17.

Focusing on the *n*-type JL-FETs first, we immediately notice from Fig. 55(a) that the raw InGaAs device variability is lower than that of Si for the metrics $V_{T,lin}$, $V_{T,sat}$, SS, and DIBL. In Fig. 57(a), if we examine the nominal band diagrams along the center of the InGaAs channel at different biases, however, we notice that the channel is only truly degenerate above threshold, whereas the source and drain extensions are always degenerate. Because the channel electron density will be significantly lower than the nominal doping, screening effects will be less important. However, at threshold, the InGaAs channel is somewhat degenerate near the top of the barrier, so the electron density will be comparatively less sensitive to small (local) changes in potential from RDF. In

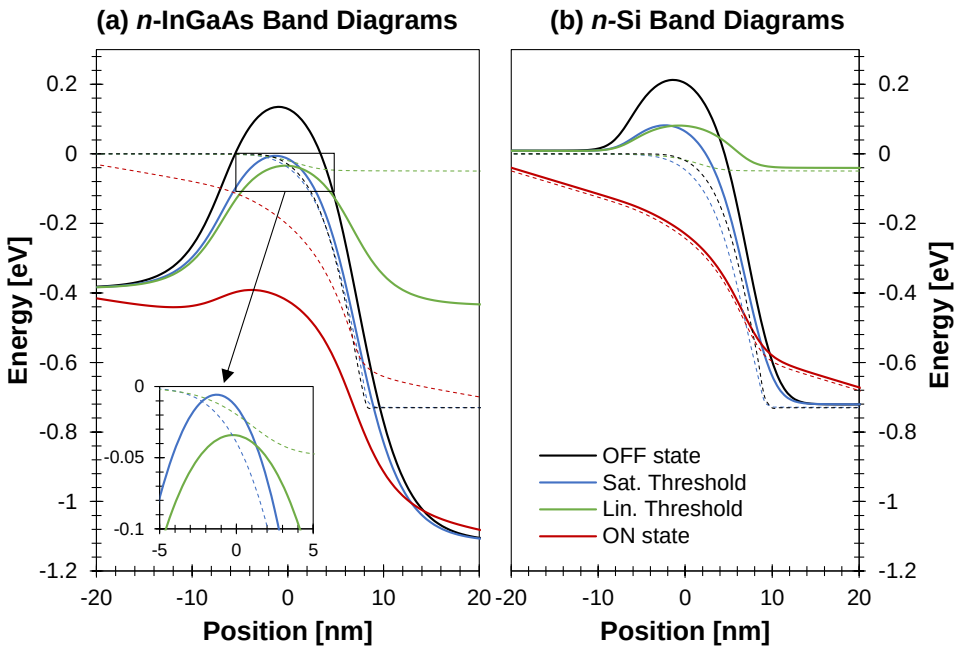


Fig. 57. Nominal conduction band diagrams along the center of the channel in *n*-type (a) InGaAs and (b) Si JL-FETs under the following bias conditions (displayed from top to bottom): off-state, saturation threshold, linear threshold, and on-state. The electron quasi-Fermi energy level is shown in dashed lines for each bias condition. The inset in (a) compares the bands at linear and saturation threshold near the top of the barrier, indicating greater degeneracy at $V_G = V_{T,lin}$ compared to $V_{T,sat}$.

other words, because of the lower DOS in InGaAs, any local alterations of the channel potential from RDF will result in smaller changes in electron density within the channel due to lower dn/dE_F as was seen in Fig. 48(b), therefore resulting in smaller threshold voltage shifts. For the Si device, the channel is always nondegenerate as shown in Fig. 57(b), so local dopant fluctuations have a stronger influence on the electron density profile, and hence, the threshold voltage.

The other factor to consider here for V_T variation is the difference in relative permittivity ϵ_r between InGaAs (13.9) and Si (11.7). From the dependence of V_T on the channel depletion width, the fluctuations in V_T can be shown to be $\sigma V_T \sim 1/\epsilon_r$ [98] so higher permittivity materials like InGaAs should have lower σV_T, all else equal. However, we will see later that the permittivity difference is only partially responsible for the lower σV_T in *n*-InGaAs when we compare the In-GaAs NFET to the PFET.

Interestingly, the difference in raw V_T variation between InGaAs and Si is more pronounced for $V_{T,lin}$ than it is for $V_{T,sat}$. The electron quasi-Fermi level in the InGaAs channel is more degenerate at linear threshold than it is in saturation threshold (see inset of Fig. 57(a)), ergo the smaller $\sigma V_{T,lin}$ compared to $\sigma V_{T,sat}$. When these results are viewed as normalized percentages in Fig. 56(a) and (b), this difference becomes even more apparent. The reduced fluctuations in $V_{T,lin}$ compared to $\sigma V_{T,sat}$ naturally lead to reduced DIBL variability in InGaAs as well.

The lower SS variations in *n*-InGaAs are connected to another electrostatic effect resulting from degeneracy. In subthreshold, the source and drain extensions remain degenerate even when the channel is not, meaning the barrier height must be higher in *n*-InGaAs compared to *n*-Si in order to reach the current criterion at which SS is extracted; this is evident from both Fig. 50 and Fig. 57 in the off state. Invoking a simple model for the FET channel potential which is derived in Appendix II of this chapter, it can be shown that there is less sensitivity of the SS to changes in barrier height for nominally larger barriers, as we see in the case of *n*-InGaAs. Essentially this means that local doping fluctuations in the source and drain regions will not modify the channel barrier as much due to stronger gate control, hence the smaller fluctuation in SS due to RDF. Moreover, if we vary the (uniform) channel doping N from 5×10^{18} cm^{-3} to 5×10^{19} cm^{-3}, we see in Fig. 58 that the sensitivity values (i.e., the slopes) of SS and DIBL to N are markedly different between *n*-InGaAs and *n*-Si near the vicinity of 2×10^{19} cm^{-3} doping. On the other hand, the SS and DIBL sensitivities are very similar for both *p*-type devices, which are in turn similar to *n*-Si as well. This is further evidence that degeneracy effects are responsible for the lower variation in subthreshold metrics for *n*-InGaAs.

The raw I_{on} variations are much larger in *n*-InGaAs than in *n*-Si (Fig. 55(c)), but this is expected given the higher nominal I_{on} in *n*-InGaAs compared to *n*-Si. In fact, the normalized I_{on}

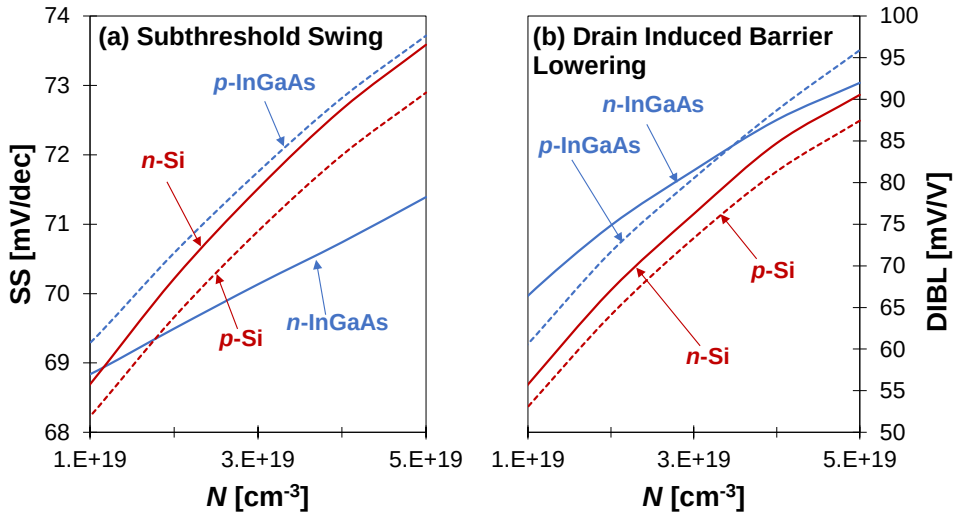


Fig. 58. (a) Dependence of SS and (b) DIBL on the nominal channel doping in Si and InGaAs JL-FETs. The sensitivity of SS and DIBL to N is lower for n-InGaAs JL-FETs due to degeneracy effects.

variations in n-InGaAs and n-Si from Fig. 56(c) are nearly identical which is similar to what we saw in the doped slab results when doping dependent mobility is ignored. Since the channel of a JL-FET in the on-state is essentially resistor-like, it is unsurprising that it has similar variability features. We must remind the reader, however, that by neglecting doping dependent mobility in our JL-FET simulations we may miss out on some of the RDF effects including greater suppression of on state current variation as witnessed in Fig. 54(c)–(d) for the Si slabs compared to InGaAs. Whether or not these effects can be extended to the case of thin JL-FET channels, however, is unknown at this point. The doubt arises from whether or not the bulk-derived doping dependent parameters are applicable to the case of ultrathin channels, and in light of the difficulty to properly calibrate a set of doping dependent parameters from NEGF, this is not a question we can answer at this time, unfortunately.

The n-InGaAs JL-FET shows a slightly larger raw σI_{off} but a smaller normalized σI_{off} when compared to n-Si. However, because of the exponential dependence of leakage current on V_{GS}, these values can fluctuate depending on sample size and the exact baseline I_{off} obtained during

calibration, so we cannot draw any precise conclusions about leakage current variation between the InGaAs and Si JL-FETs from RDF at this time.

Although the low hole mobility of InGaAs makes *p*-type devices less technologically interesting than *n*-type, the significantly higher valence band DOS makes carrier degeneracy minor in *p*-InGaAs for the same doping. Therefore, any deviations in their variability behavior from NFETs can be attributed to degeneracy effects. In Fig. 55 and Fig. 56 we see that the variability differences between *p*-InGaAs and *p*-Si are indeed smaller than what was observed for between the *n*-InGaAs and *n*-Si devices, with the remaining discrepancy among the PFETs likely stemming from the material permittivity difference.

Comparing the *n*-InGaAs and *p*-InGaAs variability results, we see that, in general, the *p*-InGaAs device is more sensitive to RDF especially for $\sigma V_{T,lin}$, σSS, and σDIBL. In this case, degeneracy is the sole reason for the lowered *n*-InGaAs variability compared to *p*-InGaAs since permittivity is identical between the two devices. To prove this, we also simulated a hypothetical scenario in which the 2×10^{19} cm^{-3} doped *n*-InGaAs JL-FETs obey Maxwell-Boltzmann statistics rather than Fermi-Dirac statistics; the results in Table 20 reveal that the *n*-InGaAs and *p*-InGaAs variability become virtually identical when degeneracy is removed.

Table 20. Differences in InGaAs JL-FET Variability Based on Carrier Model

Fermi Statistics	$\sigma V_{T,lin}$ [V]	$\sigma V_{T,sat}$ [V]	σI_{on} [A/μm]	σI_{off} [A/μm]	σSS [mV]	σDIBL [mV/V]
n-InGaAs	0.0586	0.0676	2.62×10^{-4}	3.28×10^{-6}	0.668	17.141
p-InGaAs	0.0627	0.0688	3.73×10^{-4}	4.53×10^{-6}	1.026	23.954
% Diff.	**6.8%**	**1.8%**	**34.9%**	**32.1%**	**42.2%**	**33.2%**
Boltzmann Statistics	$\sigma V_{T,lin}$ [V]	$\sigma V_{T,sat}$ [V]	σI_{on} [A/μm]	σI_{off} [A/μm]	σSS [mV]	σDIBL [mV/V]
n-InGaAs	0.0626	0.0683	3.88×10^{-4}	3.54×10^{-6}	1.110	24.857
p-InGaAs	0.0627	0.0688	3.73×10^{-4}	4.53×10^{-6}	1.026	23.954
% Diff.	**0.2%**	**0.8%**	**3.9%**	**24.7%**	**7.9%**	**3.7%**

Note: Both InGaAs JL-FETs used the exact same calibrated transport parameters, namely those for *p*-InGaAs in Table 19, for a fair comparison.

Table 21. Differences in *n*-JL-FET Variability Based on Calibration Setting

Calibration Setting	$\sigma V_{T,lin}$ [V]	$\sigma V_{T,sat}$ [V]	σI_{on} [A/μm]	σI_{off} [A/μm]	σSS [mV]	σDIBL [mV/V]
n-Si (scattering)	0.0609	0.0687	5.63×10^{-4}	6.12×10^{-6}	1.543	23.184
n-Si (ballistic)	0.0605	0.0679	2.66×10^{-4}	3.86×10^{-6}	1.686	22.815
% Diff.	**0.7%**	**1.2%**	**71.8%**	**45.3%**	**8.8%**	**1.6%**
n-InGaAs (scattering)	0.0491	0.0632	4.68×10^{-4}	3.73×10^{-6}	0.858	15.458
n-InGaAs (ballistic)	0.0495	0.0639	6.87×10^{-4}	4.49×10^{-6}	0.819	15.463
% Diff.	**0.9%**	**1.1%**	**37.8%**	**18.6%**	**4.6%**	**0.0%**
Calibration Setting	$\sigma V_{T,lin}$ (norm)	$\sigma V_{T,sat}$ (norm)	σI_{on} (norm)	σI_{off} (norm)	σSS (norm)	σDIBL (norm)
n-Si (scattering)	25.6%	47.4%	16.5%	3813.4%	2.4%	34.0%
n-Si (ballistic)	21.9%	51.3%	18.1%	5341.4%	2.2%	40.2%
% Diff.	**15.8%**	**8.0%**	**9.4%**	**33.4%**	**7.5%**	**16.6%**
n-InGaAs (scattering)	18.6%	48.5%	16.6%	2700.0%	1.2%	19.2%
n-InGaAs (ballistic)	17.6%	48.3%	15.9%	3897.7%	1.2%	21.6%
% Diff.	**5.3%**	**0.6%**	**4.6%**	**36.3%**	**4.1%**	**11.6%**

Due to the computational limitations already discussed in Section 6.3, our simulated PFETs were calibrated using ballistic data whereas the NFETs were fitted to NEGF calculations including scattering. To show that our variability conclusions remain qualitatively valid despite this discrepancy, we also performed simulations using *n*-Si and *n*-InGaAs JL-FETs calibrated against ballistic NEGF data. As shown in Table 21, the raw variations are not appreciably different between the ballistic and nonballistic cases except for the metrics σI_{on} and σI_{off}, for obvious reasons. Notably, variability of the electrostatically-driven metrics $V_{T,lin}$, $V_{T,sat}$, SS, and DIBL are not significantly affected (<10% change) by the calibration setting. When the variability results are percentage normalized, the difference in σI_{on} between the ballistic and nonballistic cases drops to below 10%, meaning that relative comparisons of σI_{on} between NFETs and PFETs remain valid despite vast differences in baseline I_{on} (as in Fig. 51(a)). These results indicate that the trends in our variability results are independent of the specific calibration results.

6.7 Summary

We compared the effects of random dopant fluctuation on equivalently designed 15nm Si and $In_{0.53}Ga_{0.47}As$ channel JL-FETs using NEGF-calibrated TCAD simulations. Degeneracy in the *n*-InGaAs device at 2×10^{19} cm^{-3} channel doping results in lower variation in threshold voltage, subthreshold swing, and drain induced barrier lowering when compared to both Si and *p*-InGaAs (which are relatively nondegenerate). For those metrics related to electrostatic device integrity (e.g., V_T, SS, and DIBL), degeneracy leads to suppressed carrier response to local potential fluctuations which enables tighter gate control of the barrier height, thus leading to smaller variations in those metrics for *n*-InGaAs. On the other hand, relative variability in on-state drive current is nearly identical for InGaAs and Si because the higher potential fluctuations also result in higher local built-in field variations which perturb current flow, despite lower fluctuations in carrier density.

Overall, *n*-InGaAs JL-FETs are more resilient to RDF due to degeneracy effects when compared to equivalent *p*-InGaAs and Si JL-FETs. Our conclusions should apply to other low DOS III-V material systems as well, thus providing additional motivation for continued research in developing heterogeneous integration technology for integrated circuits which will be the topic of the next chapter.

6.8 Appendix I: Details on NEGF Simulations and Scattering Mechanisms

Here, we provide additional details about our implementation of different scattering mechanisms in the NEGF simulations which were not covered in the main sections of this chapter.

6.8.1 Impurity Scattering

Rigorous treatment of impurities within NEGF requires either 1) simulation of an ensemble of devices with randomly placed impurity potentials or 2) the inclusion of an appropriate self-energy for the electron-impurity interaction [99]. While the former option accounts for multiple scattering and inhomogeneous screening and enables simultaneous study of device variability, it is extremely computationally expensive owing to the many simulations required as well as the need for 3-D calculations to capture the Coulomb electrostatics. The self-energy route enables a self-consistent calculation using only a single device, but its accuracy hinges on the choice of impurity potential.

The Brooks-Herring model is widely used for analyzing semiconductor impurity scattering and treats each dopant ion as a Thomas-Fermi screened Coulomb potential [100]. However, when implemented in real space NEGF simulations, the finite range of the potential leads to a nonlocal self-energy, ruling out the use of recursive Green's function algorithms and greatly increasing memory and CPU requirements [101]. Furthermore, it is well known that the Brooks-Herring model generally overestimates the mobility of heavily doped semiconductors [100]; it is still not fully understood whether this failure is due to shortcomings of the model itself or neglect of other relaxation mechanisms such as plasmon scattering [102]. These uncertainties complicate a detailed first-principles treatment of scattering in heavily doped structures.

Therefore, in our NEGF simulations we take a phenomenological approach and use an adjustable δ-function potential to derive the self-energy [103]. We assume the impurities are randomly placed and uncorrelated and only contribute to intravalley scattering, which is generally true for the Γ valley in InGaAs and is justified in bulk silicon by the large separations in momentum space between the Δ conduction band valleys. This leads to the following first-order self-energies Σ which are local at a given position $\vec{r}$:

$$\Sigma_v^i(\vec{r}, E) = N(\vec{r})\gamma^2 \int d\overrightarrow{k_\perp'} G_v^i\left(\vec{r}, E, \overrightarrow{k_\perp'}\right) \tag{6}$$

Where $N(\vec{r})$ is the local dopant density, γ is the scattering matrix element, $\overrightarrow{k_\perp'}$ is the crystal momentum, v is the valley index, and $i = ret, <, >$ for the retarded, lesser, and greater self-energies and Greens' functions. This form of the self-energy can be interpreted as a general elastic momentum-relaxing dephasing process [104]. A similar approach has been used for impurity scattering in nanowires in [105].

To fit the scattering parameter γ to experimental data, we use the Kubo-Greenwood formula [106], [107] for the mobility

$$\mu = \frac{2q}{3n}\int_0^\infty \rho(E)v^2(E)\tau(E)\frac{\partial f(E)}{\partial E_F}dE \tag{7}$$

Where n is the electron concentration, $\rho(E)$ is the density of states (which includes nonparabolicity for InGaAs) at energy E, $v(E)$ is the group velocity, $f(E)$ is the Fermi-Dirac function for Fermi energy E_F, and $\tau(E)$ is the energy-dependent relaxation time. From the Fermi golden rule we can approximate

$$\tau(E) = \frac{\hbar}{2\pi N_d \gamma^2 \rho(E)}. \tag{8}$$

Table 22. Doping-Dependent Self-Energy Parameters

N (cm^{-3})	μ_{Si} (cm^2/Vs)	γ_{Si}^2 (eV^2/cm^6)	μ_{InGaAs} (cm^2/Vs)	γ_{InGaAs}^2 (eV^2/cm^6)
5×10^{18}	143	5.39×10^{-42}	2325	1.06×10^{-41}
1×10^{19}	120	3.48×10^{-42}	1815	4.09×10^{-42}
2×10^{19}	97	2.06×10^{-42}	1422	1.48×10^{-42}
3×10^{19}	89	1.43×10^{-42}	1237	8.43×10^{-43}
4×10^{19}	85	1.09×10^{-42}	1126	5.13×10^{-43}
5×10^{19}	82	8.72×10^{-42}	842	3.61×10^{-43}

Assuming that the experimental mobility μ is limited by impurity scattering and generalizing for nonparabolic bands with nonparabolicity parameter α, we obtain

$$\gamma^2 = \frac{2q\hbar}{3\pi m^* N_d n\mu}\int_0^\infty \frac{E(1+\alpha E)}{(1+2\alpha E)^2}\frac{\partial f(E)}{\partial E_F}dE\,. \tag{9}$$

The resulting doping-dependent scattering parameters for silicon and InGaAs are given in Table 22. 1-D NEGF simulations of long resistors confirm that the calculated values reproduce experimental bulk mobilities well.

Using a fixed γ for Σ_i neglects changes in the self-consistent screening due to the inhomogeneous bias-dependent electron densities within the transistor [108], [109] as well as the surrounding dielectric environment [110]. No doubt additional physical effects due to free carrier and dielectric screening, plasmon scattering and other electron-electron effects, etc., will make contributions in real devices and lead to quantitative corrections to our numerical results, but at a prohibitive cost in computational complexity. Nonetheless, our simplified approach does allow fitting to experimental mobility data, giving it some empirical credence, and is preferable to neglecting impurity scattering altogether. Quantitatively it allows us to estimate some realistic "upper limits" on device performance and, more importantly, demonstrates that scattering effects will be qualitatively important even in ultrascaled junctionless devices.

6.8.2 Phonon and Surface Roughness Scattering

Approximately elastic interactions like low-energy phonons (i.e., long-wavelength acoustic modes) and alloy scattering (in InGaAs) are ignored in our simulations, since their contribution in heavily doped samples may be assumed to be minor compared to impurities. However, inelastic phonon scattering, particularly via intervalley optical modes in silicon and polar optical phonons (POP) in InGaAs, can lead to energy relaxation and is therefore included in the self-consistent Born approximation using standard matrix elements [111], [112]. POP is modeled as a local self-energy for computational efficiency [113].

In ultrathin-body (UTB) films, SR scattering plays an important role and is likely responsible for the experimentally observed power-law dependence of mobility on thickness in UTB silicon-on-insulator (SOI) MOSFETs [114]. Similar to the case of impurities, SR scattering can be incorporated in NEGF via simulation of an ensemble of devices with randomly generated interface roughness profiles [115] or an appropriate self-energy [116]. Again we choose the latter, adapting the model for interface roughness in [116] with an exponential autocorrelation function with correlation length λ and amplitude Δ [112] and nonlocal components approximated using adjacent diagonal elements of the Green's functions [117]. SR is included at both interfaces of the DG structure, implicitly accounting for thickness fluctuation effects [114]. Values of λ and Δ are chosen from literature fits to field-dependent mobility in experimental UTB structures; in both materials, λ is 2 nm while Δ = 0.4 nm and 1.76 nm in Si and InGaAs, respectively [112]. The large value of Δ in InGaAs is fitted to the measured mobility of wafer-bonded sub-10 nm films reported in [118].

We note that additional scattering mechanisms like interface states may be important in experimentally reported results at present, particularly in the less mature InGaAs technology [119];

we neglect these effects because it is possible that further research and development may minimize the contributions of these "extrinsic" mechanisms. SR effects, too, are basically technologically dependent, in contrast to impurity and phonon scattering, which are essentially intrinsic to doped materials; therefore, a mature process will be expected to have better optimized interfaces and hence reduced (though not necessarily negligible) scattering. As an example, note the substantial increase in experimentally reported mobility of 9 nm thick InGaAs films (about 2X) within the interval between [118] and [120]. Using, say, $\Delta = 1.76$ nm in the SR scattering model in InGaAs assumes that film quality in future III-V commercial technology will not be substantially better than that in [118]. At the opposite extreme, neglect of SR altogether corresponds to analyzing the ideal case where only "intrinsic" scattering mechanisms like impurities and POP limit transport. To cover both extremes, simulations with and without SR self-energies were presented in Section 6.3.

6.9 Appendix II: Effect of Barrier Height on Subthreshold Swing

We use a pseudo-2-D potential model to illustrate the relationship between SS and barrier height, including the *N*-dependent depletion regions in the source and drain. Such models, originally developed for IM-FETs, also apply to JL-FETs provided the electron density in the channel is small, as is generally the case in subthreshold [121]. In particular they establish the relationship between V_{GS} and the top of the barrier energy E_{TOB} in the channel and can therefore estimate SS.

$$E_{TOB} = -V_{GS} - 2\sqrt{AB} \tag{10}$$

where

$$A = \frac{-V_{GS} + V_{DS} - V_{D,dep} + \left(V_{S,dep} + V_{GS}\right) \exp\left(-\frac{L_g}{\lambda}\right)}{2 \sinh\left(\frac{L_g}{\lambda}\right)} \tag{11}$$

$$B = \frac{V_{D,dep} - V_{DS} + V_{GS} + \left(\psi_{ch} - V_{S,dep}\right) \exp\left(\frac{L_g}{\lambda}\right)}{2 \sinh\left(\frac{L_g}{\lambda}\right)} \tag{12}$$

$$V_{S,dep} = -V_{GS} + V_{S0} - \sqrt{2\psi_{ch} V_{S0} + V_{S0}^2} \tag{13}$$

$$V_{D,dep} = V_{DS} - V_{GS} + V_{S0} - \sqrt{2(V_{DS} - V_{GS}) V_{S0} + V_{S0}^2} \tag{14}$$

$$V_{S0} = \frac{q^2 N \lambda^2}{\epsilon_r \coth^2\left(\frac{L_g}{\lambda}\right)}. \tag{15}$$

In these equations λ is the DG scaling length (which depends on ϵ_r) and the effective gate voltage V_{GS} is normalized such that V_{GS} = 0 under flat band conditions. (10) can be inverted to obtain V_{GS} as a function of E_{TOB}. The subthreshold swing can then be approximated via

$$SS = \frac{\partial V_{GS}}{\partial \ln(I_D)} \sim \ln(10) \frac{\partial V_{GS}}{\partial E_{TOB}} \tag{16}$$

assuming diffusive current flow.

In Fig. 59 we plot the resulting SS as a function of barrier height for values of L_g/λ appropriate for Si and InGaAs. As expected, at the same barrier height, the SS of InGaAs is higher because of larger permittivity. However, from Fig. 50 we see that E_{TOB} ~ 0.4 eV in InGaAs and 0.2 eV in Si in the off-state leading to roughly equivalent SS in agreement with what was observed in our NEGF simulations. We also note that at higher nominal barrier heights the SS is less sensi-

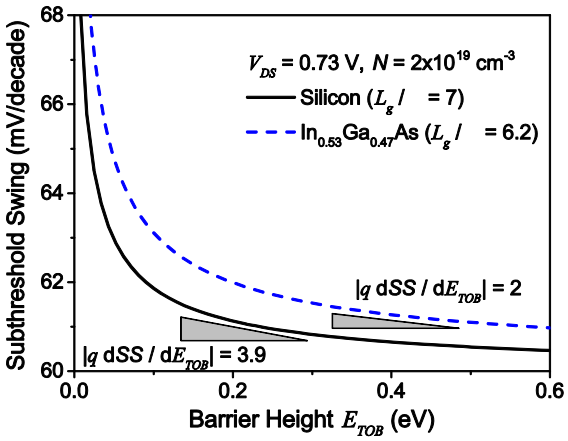


Fig. 59. Pseudo 2-D model subthreshold swing as a function of barrier height using geometric and material parameters from Table 17. The sensitivity of SS to barrier height (equal to $q\,dSS/dE_{TOB}$) is also shown at E_{TOB} = 0.2 eV and E_{TOB} = 0.4 eV for Si and InGaAs, respectively.

tive to changes in E_{TOB} because of smaller $|dSS/dE_{TOB}|$. Since these findings only rely on a generic channel potential model and the presence of strong degeneracy in the source and drain, they may also apply to other types of III-V transistors besides JL-FETs.

Chapter 7

Heterogeneous Integration Technology

7.1 Background

Heterogeneous integration (HGI) is a broad term which encompasses a wide scope of on-going research aimed to co-integrate different material systems onto a common platform to improve system performance and functionality. Every material system (e.g., Si, Ge, Group III-V, etc.) entails a unique set of advantages and disadvantages when used in electronic applications, and as a result, implementing a product with a single material technology always involves a compromise. Silicon, for example, is highly scalable and robust which has led to its commercial success in nearly all facets of the semiconductor industry over the past four decades. However, the electronic performance of Si is relatively unimpressive when compared against other materials such as Ge or III-Vs, those of which are mainly used in high-speed or optoelectronic niche applications. While such materials tend to possess higher carrier mobility than Si (recall Fig. 44 in Chapter 6), they are often plagued by quality control issues, processing challenges, and incompatibilities with standard Si CMOS foundry infrastructure. Co-integration technologies which combine the individual strengths of different material systems on a common platform (e.g., a 12” silicon wafer) may promise significant benefits in analog/RF and digital circuit applications. The key challenges involved in successful HGI demonstrations are fundamentally related to: difficulties in heterogeneous device fabrication, minimizing the HGI pitch/density, and ensuring sufficient yield and throughput.

HGI fabrication methods usually fall under one of the following categories: 1) wafer/die bonding, 2) heteroepitaxy, or 3) micro/nanotransfer printing. In the remainder of this section, we briefly discuss each of these approaches in terms of their promises and challenges.

Wafer bonding has been the preferred technique in recent HGI efforts because of the relative ease in separately processing individual wafers followed by subsequent bonding and interconnection with large vias to form packages resembling three-dimensional integrated circuits (3DICs). The main limitations of wafer bonding are: limited heterogeneous integration pitch (essentially given by the via size which is typically several μm deep), minimum alignment/overlay tolerances, wafer size mismatch of different materials, and poor yield from defects. However, the high quality of heterogeneously integrated devices and processing ease make wafer bonding an attractive solution for experimental and small-scale HGI demonstrations.

Heteroepitaxy involves directly growing heterogeneous materials during device fabrication, allowing direct in-situ integration of multiple material devices on a common platform. This method promises the highest level of HGI complexity and integration density with an interconnect pitch only limited by lithography capabilities. Unfortunately, due to lattice structure mismatch of different materials, it is difficult to grow high quality heterostructures epitaxially without introducing large dislocation densities and/or anti-phase domains, or without relying on thick buffer layers. Thermal budget concerns can also pose difficulties during processing. If these issues can be solved, heteroepitaxy-based HGI may likely offer the most significant performance benefits with minimum impact on VLSI circuit design.

Transfer-based methods generally involve physically transferring nanostructures of one material system to a receiving substrate of another material. This can be accomplished with digitized semiconductor features (e.g., wires, ribbons, fins, etc.) which are epitaxially grown on a donor

substrate and subsequently transferred to another by a "stamping" process. The main limitations of this technique are the proximity to which heterogeneous materials can be aligned and placed next to one another and uniformity control in the transfer process. These may limit the HGI pitch to values much larger (μm or higher) than what heteroepitaxy can deliver. Exotic techniques such as nanoimprint lithography and scanning probe lithography can also be used for HGI applications, but they also have limitations such as throughput and durability.

Currently, there have been a number of successful HGI demonstrations in which III-V transistors were co-integrated on a Si substrate. Many of these were sponsored by DARPA's Compound Semiconductor Materials on Silicon (COSMOS) program [123], with the objective of integrating high-speed III-V heterojunction bipolar transistors (HBTs) on a Si CMOS substrate. In [124], researchers at Northrup Grumman successfully bonded 0.25μm InP HBT "chiplets" onto a completely processed Si wafer containing 0.18μm CMOS circuitry with a heterogeneous interconnection pitch of 5 μm. A differential amplifier composed of InP HBTs and Si MOSFETs in the same circuit was created along with the design of a hybrid digital-to-analog converter (DAC) using high speed, high swing InP HBT analog blocks and Si CMOS digital correction blocks. These results show the possibility of transistor-level integration for analog/mixed-signal applications. In another demonstration, wafer-level packaging was used to co-integrate an antenna atop an RF front end module in a single monolithic microwave integrated circuit (MMIC) package, promising low cost, high performance, compact, hermetically sealed RF electronics [125].

Wafer bonding has also been used to cointegrate III-V and Ge transistors on Si substrates for digital applications as well. In [126], researchers from IBM successfully cointegrated InGaAs NFETs and SiGe PFETs by bonding epitaxially-grown InGaAs/InP wafers with SiGe-on-insulator

wafers to form “hybrid ETXOI” substrates, followed by downstream processing to form HGI inverters. Because entire sheets of heterogeneous material were “transferred” by wafer bonding and subsequently patterned into their corresponding N and P active regions, the N-P separation was lithographically set (at 250 nm) and not limited by overlay accuracy. In a similar technique, researchers from AIST cointegrated InGaAs NFETs and Ge PFETs by layer transfer of InGaAs/InP on germanium-on-insulator (GeOI) wafers through direct wafer bonding [127].

A number of recent works have demonstrated epitaxial growth of high quality III-V device layers on Si as well. One example is an AlGaAs/GaAs HBT [128] fabricated on SiGe/Si substrate using a graded SiGe buffer to exploit the nearly identical lattice constant between GaAs and Ge. To avoid using a dedicated buffer layer, aspect ratio trapping (ART) can be employed to prevent threading dislocations from propagating upward into the active layer during molecular beam epitaxy (MBE), followed by epitaxial lateral overgrowth (ELO) to form uniform, high quality GaAs films. In [129], GaAs MOSFETs were fabricated directly on Si using metal-organic chemical vapor deposition (MOCVD) with ART inside high aspect ratio SiO_2 trenches.

Transfer methods have also shown promise for HGI implementations. In [130] and [131], arrays of InAs nanoribbons (NRs) were epitaxially grown and patterned on an AlGaSb substrate and transferred to a SiO_2/Si substrate to ultimately form compound semiconductor (X)-OI transistors with $I_{on}/I_{off} > 10^4$. In [132], arrays of InAs and InGaSb NRs were sequentially transferred to a SiO_2/Si substrate to form XOI CMOS circuits, demonstrating true feature-level HGI. Heterogeneous integration of epitaxially-grown GaAs and Si NW arrays via polydimethylsiloxane (PDMS) stamp transfer on SiO_2 was also demonstrated from our research group [133], but with a limited heterogeneous integration pitch of ~80 μm. Researchers at the University of Illinois at Urbana-

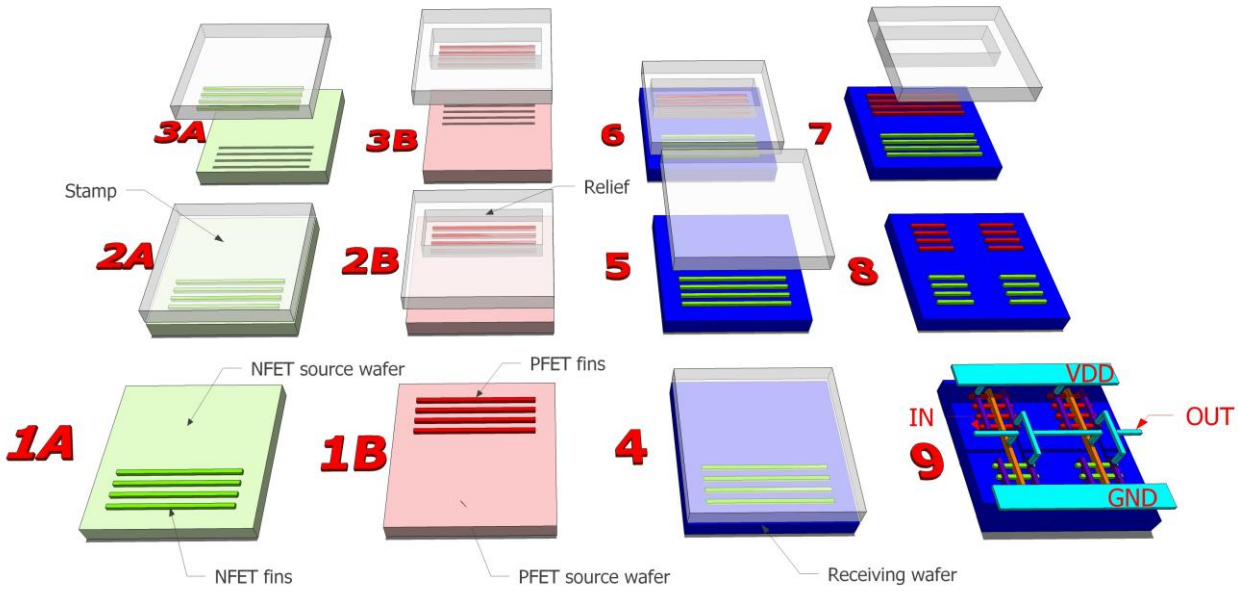
Champaign have shown numerous successful demonstrations of transfer printed devices on both rigid and flexible substrates [134], [135].

Clearly, HGI has garnered much attention from technological perspectives. Still missing, however, is a basic evaluation of the potential benefits in terms of speed, power, and area enabled by HGI technology over homogeneous (i.e., Si-only) CMOS for near-future generations. Without such knowledge, it will be difficult for the semiconductor industry to assess the true value of HGI as an alternative way to ensure continued performance gains in next-generation electronics beyond the inevitable scaling limits of Si.

Our objective in this chapter is two-fold: first, we develop a fabrication process based on the concept of nanotransfer printing (NTP) which serves to enable feature-level (e.g., transistor-to-transistor) HGI of III-V FETs on Si substrates; then, we develop a performance and cost evaluation framework to assess the potential benefits of feature-level HGI in nanoscale VLSI circuits with consideration of NTP-related penalties in actual designs. In Section 7.2, we discuss the general scheme for implementing heterogeneous circuits using NTP. In Section 7.3, we show our experimental progress on transferring GaAs and InAs nanoribbons to Si substrates and highlight key challenges of the NTP process including alignment accuracy and transfer yield. In Section 7.4, we introduce the proposed evaluation framework and apply it to the case of 15nm InGaAs/Ge FinFETs compared to Si-only FinFET technology. In Section 7.5, we compare estimated manufacturing costs for NTP-based HGI at the 15nm node against alternative HGI and non-HGI technology options. Finally, in Section 7.6 we conclude by summarizing the most important findings from this chapter.

7.2 Nanotransfer HGI Process: Proposed Concept

Here we introduce a general approach for implementing heterogeneous circuits for use in micro and nanoelectronic applications. The technique relies on nanotransfer printing to pick up and transfer digitized features from one or more “source” or “donor” substrates to a final “receiving” substrate which serves as a common platform for the cointegrated materials. The scheme is conceptually illustrated in Fig. 60 and shows how a simple FinFET buffer can be implemented using different materials for the NFET and PFET devices. The active layers are first patterned into a discrete number of fins (as an example) on their respective source wafers (Step 1). The fins are then undercut by a selective etching step which removes the underlying sacrificial layer, possibly even suspending the fins. After undercutting, an elastomeric stamp is pressed on the source substrate, causing the fins to adhere to the stamp surface (Step 2). The stamp is then released from the



1A/1B: Pattern source wafers into discrete fins
2A/2B: Pick-up N/P fins on stamp
3A/3B: Release stamp
4: Transfer N fins to receiving wafer
5: Release stamp
6: Align/transfer P fins to destination wafer
7: Release stamp
8: Trim (etch) fins between different FETs
9: Gate stack formation, doping/annealing interconnect formation, back-end metallization, etc.

Fig. 60. Process flow sequence for NTP-based HGI.

source wafer, picking up the fins because of a favorable surface energy profile at the stamp surface (Step 3). Note that the same sequence of steps is performed for each source material to be transferred. Once the stamp has picked up the NFET fins (Step 3A), it is pressed against the receiving wafer (Step 4), transferring the fins to the wafer. This transfer process again relies on a favorable surface energy profile between the fins and the receiving wafer over the stamp. Upon stamp release (Step 5), the NFET fins are successfully transferred while, in principle, preserving their original pitch, size, and number.

After the NFET fins transfer, another stamp containing PFET fins (Step 3B) is then carefully aligned and transferred to the receiving wafer (Steps 6 and 7) in a similar fashion. The alignment step is critical because it directly sets the HGI proximity and determines whether feature-level integration is possible without a significant area or yield penalty due to overlay errors. After the PFET transfer, a trim mask is used to etch away the NFET and PFET fin regions that bridge different transistors or logic gates in the circuit layout (Step 8). The use of large-area fin transfer followed by trimming has a significant benefit over small-area fin transfer for reasons to be discussed in Section 7.4. Finally, remaining process steps such as transistor gate stack formation, doping and annealing, local interconnect formation, and metallization are performed as needed and can be tailored to the process requirements for the actual integrated materials.

In general, co-integration of different materials may entail different thermal budget restrictions in downstream process steps. For example, the traditionally high temperatures ($T \geq$ 1000°C) reached during rapid thermal annealing (RTA) in Si processing may approach or even exceed the melting point for other semiconductors like InGaAs ($T_m \cong$ 1100°C) and Ge ($T_m \cong$ 938°C), while a lower temperature anneal may result in sub-optimal dopant activation for IM-FinFETs. There is evidence that Si^+ implanted *n*-InGaAs can reach near 100% activation for a 10

sec RTA between 750–850°C for electron sheet densities up to 5×10^{14} cm^{-2} [137], while B^+ implanted *p*-Ge can be fully activated even without any post-implant annealing for hole sheet densities up to 10^{14} cm^{-3} and BF_2^+ implanted *p*-Ge can be fully activated after a 30 min. low temperature anneal of 350°C [138]. Experimental demonstrations have also shown successful use of sub-800°C RTAs for post-implant dopant activation in InGaAs FETs [139]–[141] and sub-400°C fabrication of entire Ge PFETs [142]. These findings suggest that simultaneous HGI processing of InGaAs and Ge may be possible for inversion-mode devices requiring precise junction definition. On the other hand, co-integration of Ge or InGaAs with Si may be more problematic because of the much higher anneal temperatures required for dopant activation in Si.

Alternatively, uniformly doped JL-FETs are particularly suitable for HGI because of their relaxed thermal budget requirements. Since the channel materials can be doped *in situ* during growth on the source wafers, one may avoid subsequent high temperature processing such as post-implant RTA which may be especially problematic in multi-material settings. For these reasons, our experimental work to be presented in the next section is largely devoted to the transfer of uniform, heavily doped III-V nanoribbons which are suited for JL-FET applications.

In order for NTP to truly become a viable HGI solution, ultimately the process must be scalable enough to achieve high volume production in commercial foundries. This is a no simple feat and, so far, no feature-level HGI process[8] has shown such capability, although many show promise that it may be achieved one day. Besides the low thermal budget demands for NTP, one of the potential advantages of transfer-based HGI over wafer bonding is the ability to reconcile

[8] Although bonding is a well-established method used by industry, it does not provide feature-level HGI at the scale we are interested in.

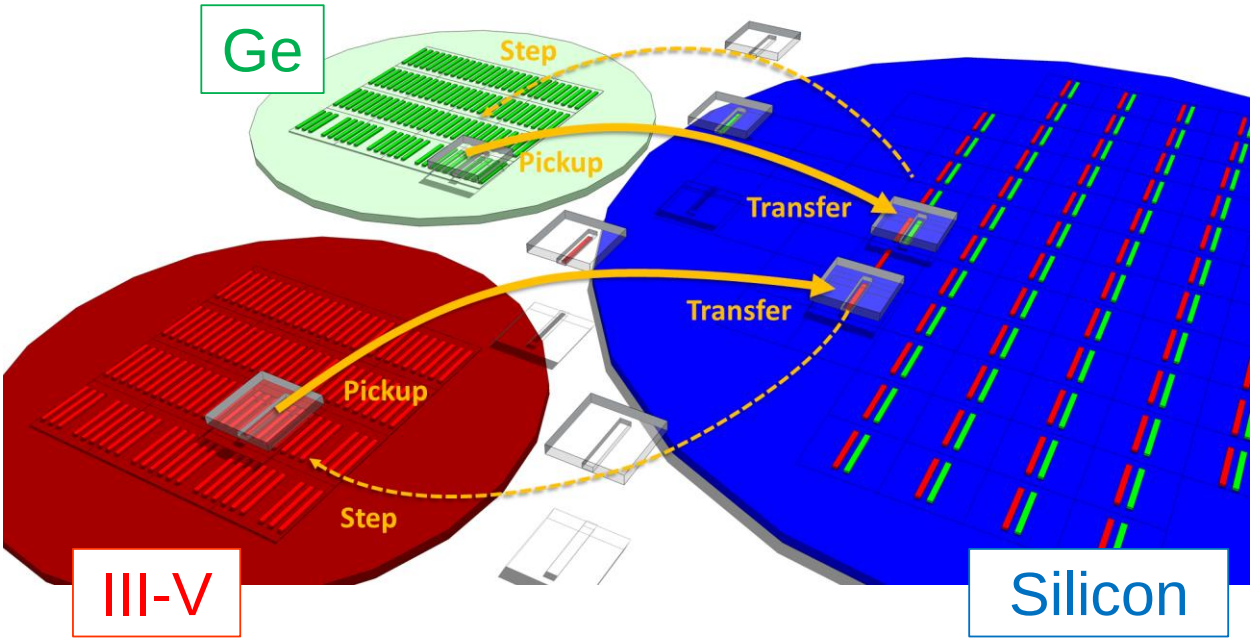


Fig. 61. Possible wafer-scalable concept of III-V/Ge HGI on Si realized through a repeatable "step and transfer" NTP process.

wafer-size mismatches through repeated transfer steps and substrate recycling. This is conceptually shown in Fig. 61 where active materials (e.g., fins, wires, ribbons, etc.) are sequentially transferred from smaller source wafers (e.g., III-V and Ge) to fully populate a larger Si wafer. If the source/donor wafers are "over-patterned" to contain more active features per die area than needed on the Si wafer, then a single donor layer could suffice for one or more fully processed Si wafers despite the size mismatch. Moreover, a superlattice of repeating active/sacrificial layers grown on the source wafer would allow continued use of the wafer even after the topmost layer material becomes exhausted during transfer, thereby allowing efficient reuse and recycling of the expensive source wafers. The repeating transfer process could be implemented using modified versions of foundry-standard "step and imprint" nanolithography tools [143], thereby realizing "step and transfer" based NTP for HGI applications. The true feasibility of such a solution is, of course, entirely speculative at this time; however it is more than likely that any real-world implementation

of NTP-based HGI would inevitably require this level of process scalability and automation to be technologically worthwhile.

7.3 Nanotransfer HGI Process: Experimental Work

7.3.1 Previous Work: Integration of GaAs NR Arrays with Si on Si/SiO_2

Our experiments to be detailed in Sections 7.3.2 and 7.3.3 are based on a continuation of the work[9] from [133] where arrays of 400 nm wide and 40 μm long GaAs nanoribbons (NRs) were transferred from a GaAs/AlGaAs/GaAs substrate to a Si/SiO_2 substrate which contained pre-patterned Si NRs of the same dimensions. The results of that effort are shown in Fig. 62(a) where we can see periodic NR arrays of GaAs and Si in close proximity (~100 μm) of each other and covering a wide area of several mm^2. We should note that the alignment of the GaAs arrays next to the Si arrays is deliberate and embodies the principles of feature-level HGI, albeit at microscale as opposed to nanoscale dimensions. In Fig. 62(b), two of the heterogeneous arrays are shown connected by evaporated metal interconnects which were patterned via optical lithography in a Karl Suss MA6 contact aligner. We should emphasize that the alignment, transfer, and metallization steps were all performed using conventional lithography with predetermined electrode layouts and that no "freehand" electron beam lithography was relied upon for interconnect formation, unlike the demonstration in [132]; this is an important distinction because, in practical settings, the metal interconnect layouts are fixed at the time of mask design and cannot arbitrarily change in response to where the transferred features actually land.

[9] The experimental results presented here are from my colleagues Jorge Kina, Dr. Kun-Huan Shih and Dr. Kyeong-Sik Shin.

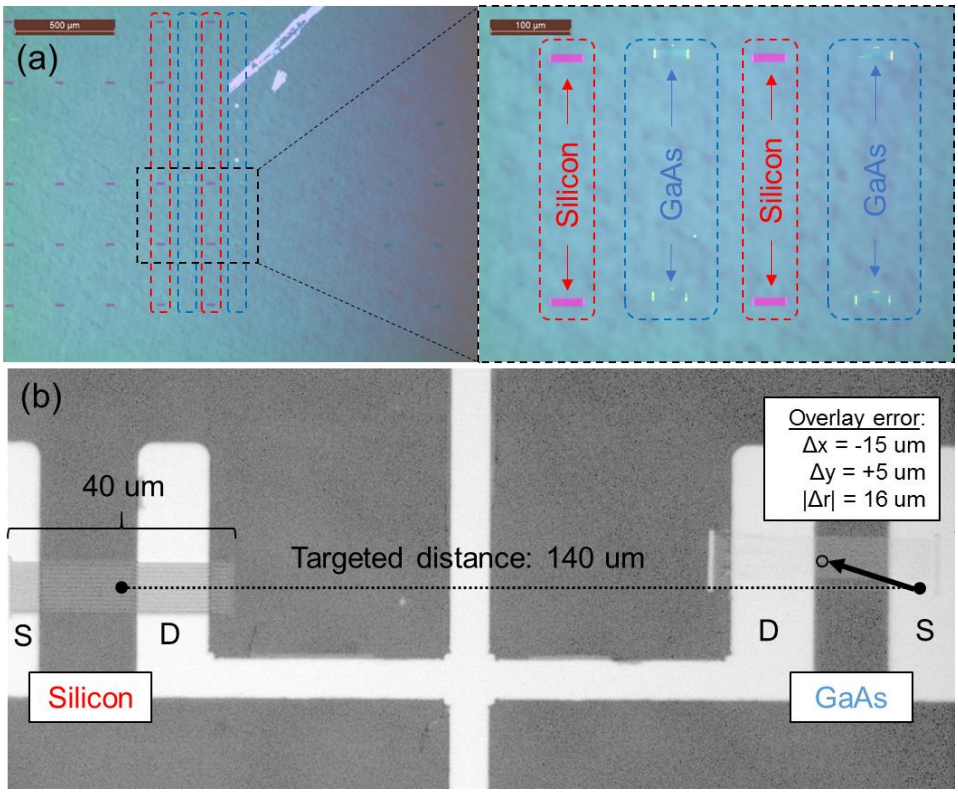


Fig. 62. (a) HGI demonstration of 400 nm wide GaAs and Si nanoribbon arrays formed by NTP on SiO_2/Si substrate with mm^2 area coverage. (b) Measured overlay error (16 μm) after aligned transfer and source/drain electrode formation using optical lithography.

As we have mentioned, the accuracy of the aligned transfer step is a critical factor for realizing NTP-based heterogeneous circuits. The primary bottlenecks to alignment accuracy are the limited resolution of the optical systems (e.g., contact or stepper aligners) used to perform the alignment, the precision of the (*x*, *y*, and θ axis) stage movement, and the topography of the stamp and receiving wafer over large areas. In Fig. 62(b) the alignment overlay error[10] for the transferred GaAs arrays with respect to the Si arrays was about 16 μm. This error is obviously too high for use in nanoscale circuits where the typical separation between NFETs and PFET may be below 100 nm. Commercial steppers [144] with overlay errors of less than 10 nm may provide the needed alignment accuracy if they can also be modified to perform the transfer process (i.e., realizing

[10] The laboratory tools available to us limit the achievable transfer accuracy to the order of several microns at best.

"step and transfer" systems), although the overlay tolerance may still exceed several tens of nm due to more severe topography issues and mechanical properties of the stamp (which is quite soft and flexible in the case of PDMS). We will revisit some of these issues later in Section 7.4.3.

7.3.2 Experiment #1: Transfer of High Aspect Ratio GaAs NR Arrays to Si/SiO₂

Continuing from the work just described, our first experimental goal is to determine whether very high aspect ratio[11] (AR) features can be successfully transferred with good yield, and if the transfer yield has any clear dependence on the feature dimensions and design. We will show in Section 7.4 that the ability to transfer high AR features will be a crucial factor in determining whether or not NTP-based HGI can offer real performance advantages in light of higher area penalties stemming from overlay errors. For this experiment, we attempt to transfer very dense arrays of 30 nm thick GaAs NRs with considerably high AR to SiO_2/Si substrates for *n*-type JL-FET applications. The complete process flow given in Appendix I at the end of this chapter.

Fig. 63 depicts the epitaxial layer stack for the GaAs source substrate. The topmost layer is the active GaAs layer with *n*-type doping of 10^{18} cm^{-3}; this value was chosen to optimize the balance between channel resistivity and maximum depletion extent for a gate overdrive of 1 V.

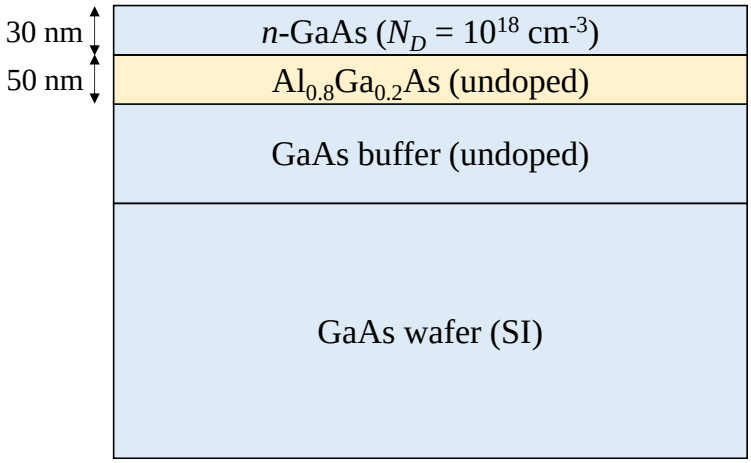


Fig. 63. MBE-grown layer stack for *n*-GaAs/$Al_{0.8}Ga_{0.2}As$/GaAs substrate.

[11] We are primarily interested in the length-to-width ratio.

Directly underneath the GaAs active layer is a 50 nm $Al_{0.8}Ga_{0.2}As$ sacrificial layer which has nearly complete etch selectivity to GaAs under aqueous citric acid/hydrogen peroxide (20:1) solution [145] and vice versa under dilute buffered oxide etch (BOE) solution [146]. A complete etch selectivity will allow us to cleanly undercut and/or release the active GaAs layer by several hundred nm prior to retrieval by the PDMS stamp.

Starting with the epitaxial GaAs/AlGaAs wafer, positive photoresist was spin-coated followed by optical lithography in an ASML PAS 5500 stepper to define the NR arrays. The arrays ranged from 25 μm to 400 μm in length while the nominal width was fixed to 0.5 μm. After lithography, the wafer was diced into smaller individual substrates for subsequent processing. To prepare each GaAs substrate for patterning, the native oxide was first removed in a mixture of 37% HCl/H_2O in a 1:5 ratio for 30 sec. Then, a solution of citric acid/hydrogen peroxide was prepared by first mixing 20 g of anhydrous citric acid with 20 mL H_2O in a 40°C heated bath with stirring, followed by the addition of 1 mL H_2O_2. The active GaAs layer was then etched in the citric acid/hydrogen peroxide solution to stop on the underlying $Al_{0.8}Ga_{0.2}As$, after which the photoresist was

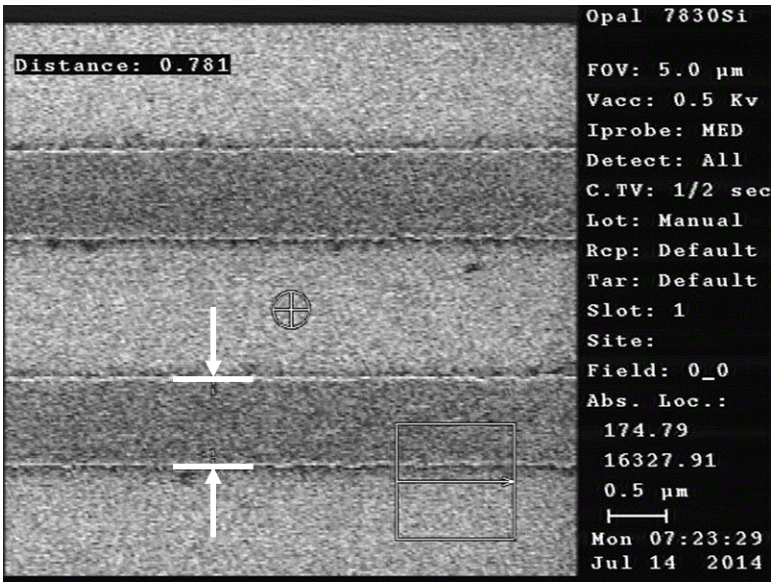


Fig. 64. CD-SEM image of as-etched GaAs NRs in citric acid/hydrogen peroxide solution. The nominal width of each ribbon is 0.5 μm, whereas the actual measured width is 0.781 μm.

removed in acetone. Once patterned, the NRs were between 0.7–0.8 μm wide as revealed by CD-SEM in Fig. 64; deviations between the actual width and the nominal value of 0.5 μm are the result of autofocus errors in the stepper and underdevelopment of the resist. The GaAs NRs were then undercut by carefully timed etching of the AlGaAs in dilute BOE solution until one or more visual indications of sufficient undercutting were observed. Once the NRs were determined to be sufficiently undercut, a PDMS stamp was gently placed in contact with the substrate and then quickly removed to pick up the NRs. The success of NR retrieval critically depends on achieving enough undercut (but not too much) while also ensuring uniformity of the undercut process across the substrate.

Fig. 65 depicts the evolution of the undercutting process: after roughly 6 – 8 min the NRs show blue highlights around the edges indicating a sufficient undercut, but after 9 min many NR segments turned translucent indicating ribbon collapse due to excessive undercutting. In general, we found that wet etching of the sacrificial layer was highly nonuniform and pattern-dependent which made it difficult to control the amount of undercutting. This presented a significant chal-

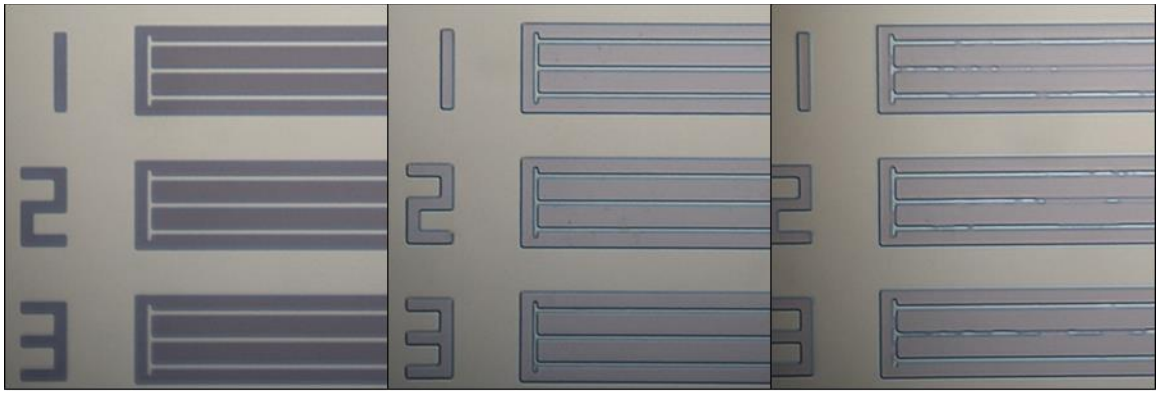


Fig. 65. Optical micrographs of GaAs NR arrays being undercut by selective etching of the underlying $Al_{0.8}Ga_{0.2}As$ after (a) 0, (b) 6, and (c) 9 min in dilute BOE solution. White areas correspond to the top layer *n*-GaAs while violet corresponds to the bottom GaAs layer. After a 9 min undercut, some NRs began collapsing as indicated by translucent segments at random locations.

lenge for several reasons. First, an insufficient undercut prevents the PDMS stamp from fully picking up the NRs which can lead to ribbon fracture during retrieval or simply no pickup at all. Second, an over-excessive undercut results in suspension of the NRs which leads to bending and/or collapse from stiction during drying, even in low surface tension isopropanol (IPA) solvent. It was observed that collapsed ribbons could not be picked up at all by PDMS due to strong adhesion of the NRs with the bottom layer GaAs. Third, we found that adding extra undercut time after a failed pickup attempt simply resulted in the GaAs substrate becoming extremely hydrophobic after initial contact with PDMS, thereby preventing the BOE etchant from working after the first few attempts.

After pickup, the PDMS stamp was dipped in BOE again for 60 sec to remove any remaining AlGaAs on the underside of the ribbons. Then, both the stamp containing NRs and the receiving SiO_2 substrate were dipped in hydrogen peroxide to form hydrophilic surfaces in preparation for transfer. The stamp was then gently pressed against the SiO_2 substrate and slowly removed to complete the NR transfer. Afterwards, the SiO_2 substrate was exposed to oxygen plasma for 60 sec at 80 W power to turn the now-hydrophobic surface after transfer to hydrophilic again and also clean the surface of any PDMS residue.

Fig. 66 shows dense arrays of 400 μm long GaAs NRs successfully transferred to a SiO_2/Si substrate using this process. Unfortunately not all of the ribbons were transferred fully intact as missing segments are clearly observed, thereby resulting in <100% transfer yield. The causes of imperfect yield are due to: 1) nonuniformities in the undercutting stage, 2) weak adhesion of PDMS to GaAs during pickup, 3) weak adhesion of GaAs to SiO_2 compared to PDMS during transfer, and 4) random probability of mechanical fracture from bending/peeling stresses during pickup and

Fig. 66. A set of 10^{18} cm^{-3} *n*-doped GaAs nanoribbon (L/W/T = 400/0.75/0.03 μm) arrays transferred to SiO_2/Si. Each individual array is nominally composed of ten parallel ribbons. Discontinuities along the nanoribbons indicate broken segments resulting in <100% yield.

transfer. In theory, most of these challenges could be addressed by optimizing the process to either minimize sources of variation or decouple their effect on the end result.

For example, using HF vapor instead of wet etching would prevent stiction-induced NR collapse during drying and allow us to fully undercut the NRs (even suspending them) without relying on precisely timed undercuts which are highly vulnerable to etch rate variations. In addition, fully suspending the NRs (except near anchors at the ribbon ends) would likely maximize the probability of successful pickup by PDMS. Unfortunately, our attempts to use HF vapor were unsuccessful: we were not only unable to pick up any NRs but we also observed etching byproducts which remained on the substrate and could not be removed without washing in water, which defeats the purpose of dry etching. We have also tried serial rinsing of wet etched (suspended) ribbons in water, 1:1 water/IPA and IPA to prevent stiction-induced collapse, however we observed NRs bending and collapsing even while still wet, possibly indicating that the NRs were simply too thin to begin with to hold structural integrity during etching. Researchers from U. Illinois at Urbana-Champaign [147] have demonstrated complete suspension and pickup/transfer of

GaAs microribbons using aqueous HF/ethanol solution, but for much wider and thicker ribbons (100 μm and 270 nm, respectively) compared to ours (0.75 μm and 30 nm, respectively).

Optimizing the surface interactions between GaAs, PDMS, and SiO_2 before the pickup and transfer stages may also increase the process yield. Since PDMS is used for both pickup and transfer, it is necessary to ensure that the adhesion strength between PDMS and GaAs is high during pickup, but weaker than that between GaAs and SiO_2 during transfer. Normally, the viscoelastic property of PDMS enables this dual property by virtue of its peel speed-dependent interfacial adhesion strength [148]. It is known that a fast peel rate (e.g., >10 cm/s) is beneficial for picking up features from the donor substrate while a slow peel rate (e.g., <1 mm/s) is beneficial for transferring features from PDMS to the receiving substrate. These guidelines were applied in our experiments, however the pickup yields were still relatively poor. To increase the interfacial adhesion strength between GaAs and PDMS, the authors in [147] deposited a thin layer of SiO_2 on top of GaAs and exposed the PDMS to ultraviolet induced ozone to form a hydroxyl terminated surface, essentially transforming the weak van der Waals interaction between GaAs and PDMS into strong covalent siloxane –Si–O–Si– bonds between GaAs and PDMS. The disadvantage of this approach is that transfer to a different substrate will become more difficult. To increase the transfer yield, one can reduce the fractional surface area of PDMS that contacts the GaAs by using perpendicular-oriented grated reliefs on the PDMS surface [149], [150]; this, however, has the opposite challenge of reducing pickup yield. Clearly, the tradeoff between pickup and transfer yield must be optimized according to the particular process and measures should be taken to improve one or the other depending on whichever step poses the largest bottleneck to yield.

Once the NRs were transferred to SiO_2, we deposited 5 nm of Al_2O_3 by atomic layer deposition (ALD) in a Cambridge NanoTech Fiji F200 chamber to serve as the gate dielectric. Afterwards, gate lithography was performed followed by evaporation of 200 nm Al at room temperature in a CHA Mark 40 e-beam evaporator and metal liftoff in acetone. Source/drain lithography was performed next followed by contact window opening in dilute BOE for 30 sec to remove Al_2O_3 covering the source/drain regions. Next, we evaporated a stack of AuGe/Ni/Au (50/25/50 nm) at room temperature followed by metal liftoff in acetone. Finally, we performed RTA at 400°C for 60 sec, completing the process flow.

An example of a completely processed GaAs JL-FET on SiO_2 is shown in Fig. 67. The device has a gate length of 7.5 μm with roughly the same amount of underlap on both sides. Besides imperfect transfer yield, there were several problems encountered during processing. During S/D contact opening, we observed that many NRs had bent or broken off from the contact regions probably due to undesired etching of the underlying oxide by HF. Also, immediately after RTA

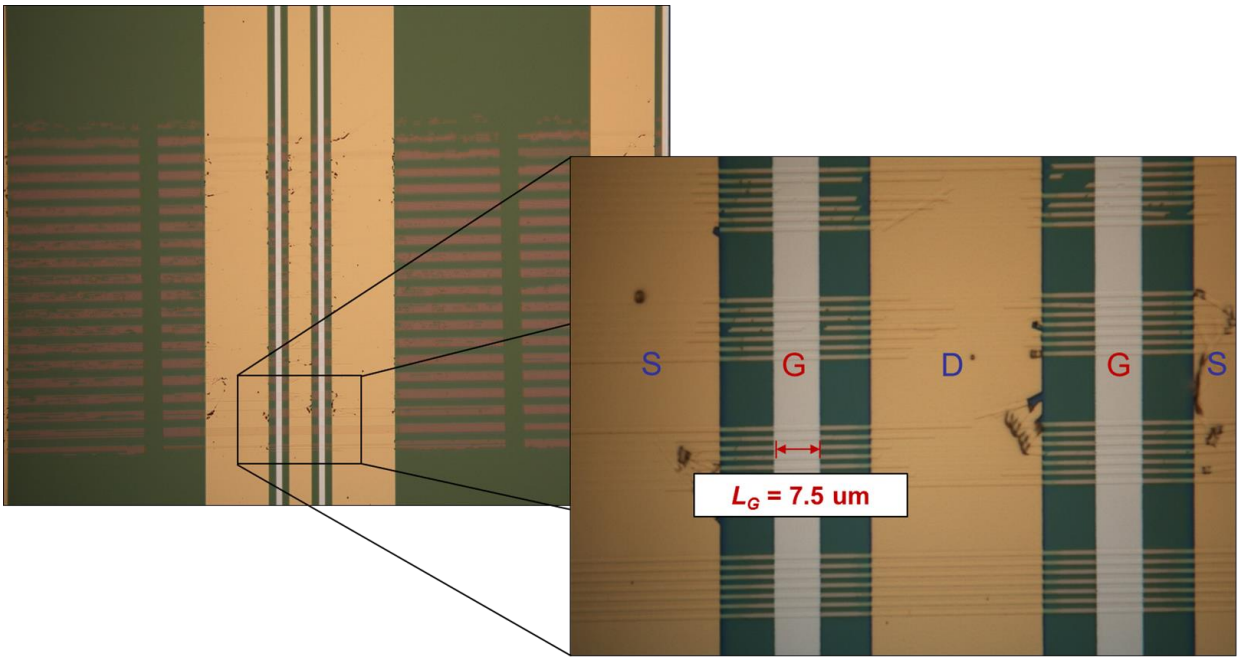


Fig. 67. Optical micrographs of a completely processed L_G = 7.5 μm *n*-GaAs JL-FET on SiO_2 substrate.

we noticed that the GaAs NRs became discolored and/or vanished near the edges of the S/D lines; the exact reason for this was never determined. Unfortunately, none of the finished JL-FETs showed signs of electrical conduction when we attempted to characterize them as the measured current level was in the picoamp or less range which is on the order of the noise floor of the parameter analyzer. Both two-terminal and three-terminal I-V measurements were carried out in ungated, front-gated, and back-gated configurations over a wide range of voltage (up to ±40 V), but no meaningful behavior could be established from the acquired data. We hypothesize that the transferred NRs were either: 1) severed during RTA, 2) had microscopic cracks or other discontinuities across their width which were not easily visible under optical inspection, 3) suffered from extremely high S/D contact resistance (perhaps due to incomplete Al_2O_3 removal in BOE or incomplete alloying of the evaporated AuGe on GaAs), or 4) otherwise materially compromised by the PDMS during transfer. Due to financial and resource constraints, we were unable to troubleshoot the individual process steps in detail or identify the exact reason(s) for device failure.

7.3.3 Experiment #2: Transfer of High Aspect Ratio InAs NR Arrays to Si/SiO_2

Our second experiment involved the transfer of large InAs NR arrays to SiO_2 to form InAs JL-FETs. InAs was explored in hopes of avoiding contact resistance problems due to the smaller band gap of 0.36 eV compared to 1.42 eV for GaAs. In addition, it is known that metal-InAs interfaces often have a Fermi level which is pinned deep in the conduction band which makes the formation of ohmic contacts to *n*-InAs even easier. The epitaxial InAs source substrate is shown in Fig. 68. Because of the higher lattice constant mismatch in InAs/GaSb (0.66%) compared to GaAs/AlAs (0.14%), the active *n*-InAs layer (10^{18} cm^{-3} doped) is made 15 nm thick which is below the critical layer thickness of ~20 nm [153]. $Al_{0.4}Ga_{0.6}Sb$ serves as the sacrificial layer in this sys-

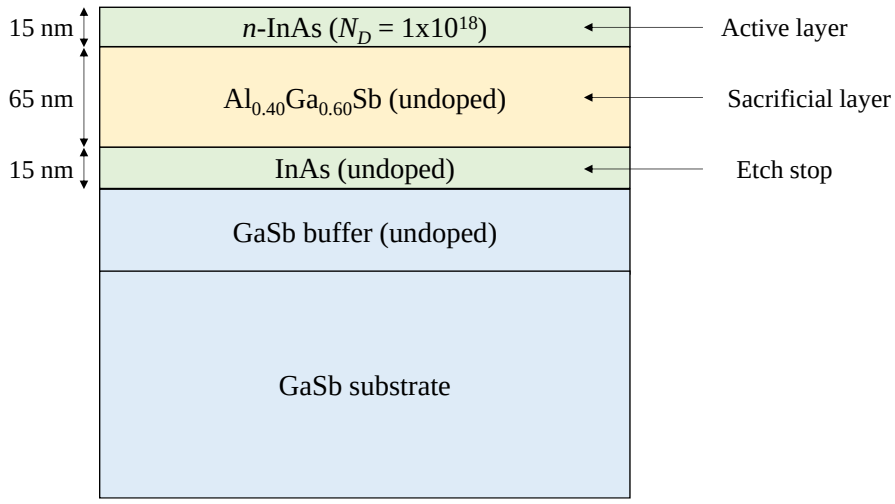


Fig. 68. MBE-grown layer stack for *n*-InAs/$Al_{0.4}Ga_{0.6}Sb$/GaSb substrate.

tem and can be selectively etched against InAs by dilute NH_4OH solution [151]. Citric acid/hydrogen peroxide solution selectively etches InAs and is used here again [152]. A second InAs layer is introduced below the sacrificial AlGaSb layer and serves as an etch stop to prevent removal of the GaSb substrate.

The complete InAs process flow is provided in Appendix II at the end of this chapter. Many of the steps are similar to those used for the GaAs process so we will not cover them in detail except to point out any major differences. Electron beam lithography was used to pattern the InAs instead of optical lithography for this experiment to test different pattern layouts. The width of each NR was set to 0.5 μm while the length varied from 100–400 μm. We found that large patterns systematically showed evidence of nonuniform etching: regions near the outer perimeter exhibiting a faster etch rate than those near the center. When undercutting the NRs in dilute NH_4OH, the etch "front" gradually progressed from the outer perimeter and eventually converged toward the center in a ring-like fashion as shown in Fig. 69(a)–(c). Because only the NR segments near the etch front have received enough undercut for pickup and transfer, but not so much as to induce NR

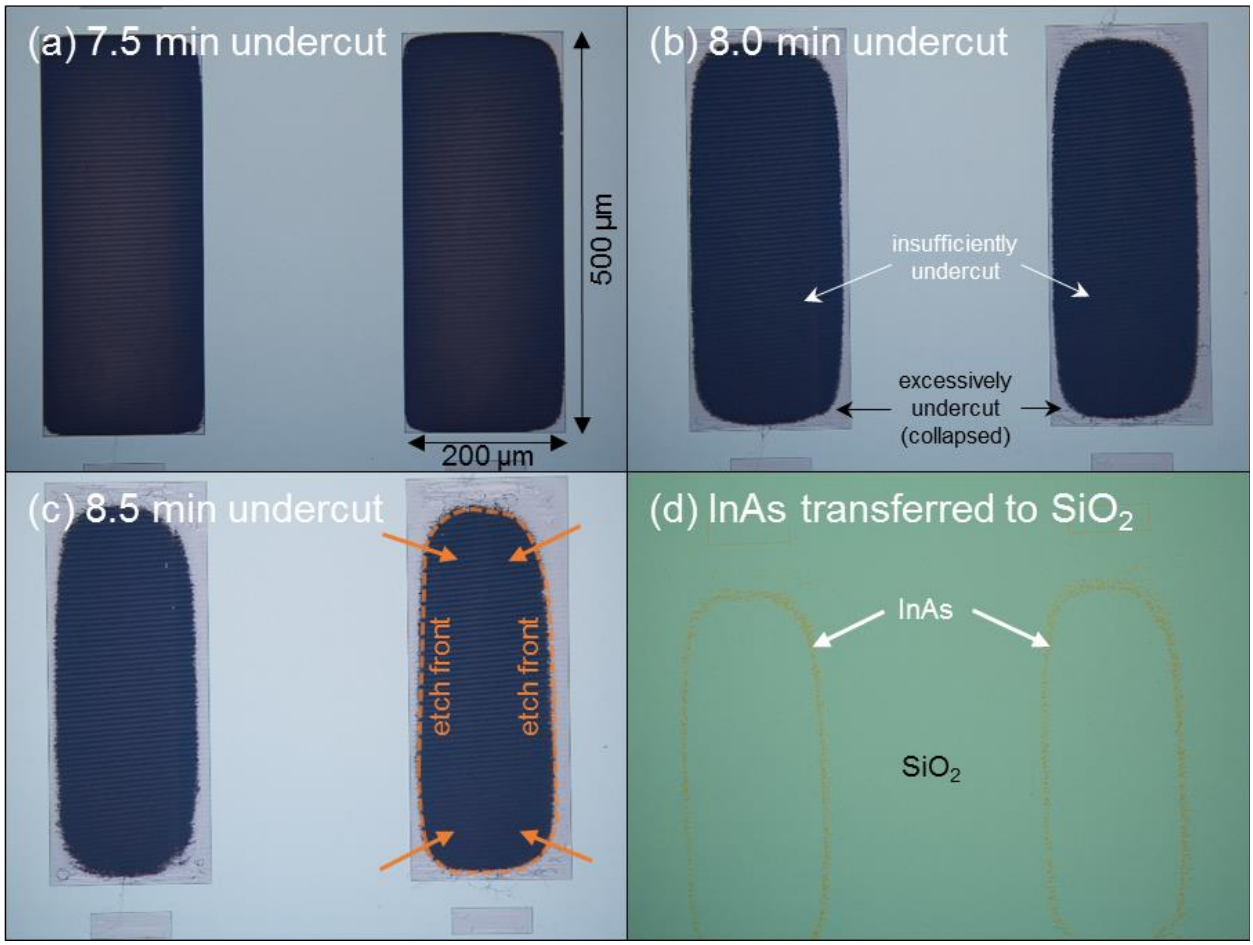


Fig. 69. Time progression of InAs NR undercutting by AlGaSb etching after (a) 7.5 min, (b), 8.0 min, and (c) 8.5 min in dilute NH_4OH solution. Each 200×500 µm^2 rectangular area contains an array of 500 parallel NRs. The dark blue regions correspond to NRs that are insufficiently undercut, while the lavender regions correspond to bent and/or collapsed NRs that received an excessive undercut. The etch front illustrated by the dashed orange outline in (c) corresponds to NR portions that are on the verge of collapse. In (d), broken portions of InAs from the etch front were successfully transferred to SiO_2, but nothing else.

collapse, the outcome of any transfer attempt resulted in an unusable broken ring pattern as depicted in Fig. 69(d). We should point out that similar results were obtained from our failed GaAs transfer attempts as well, so this problem was not unique to the InAs process. We hypothesized that this problem was the result of improper substrate wetting prior to etching, but despite several attempts to improve the wetting and etch uniformity, no significant improvement could be made. Instead, we found that smaller and more isolated sets of NR arrays showed more uniform etching rates.

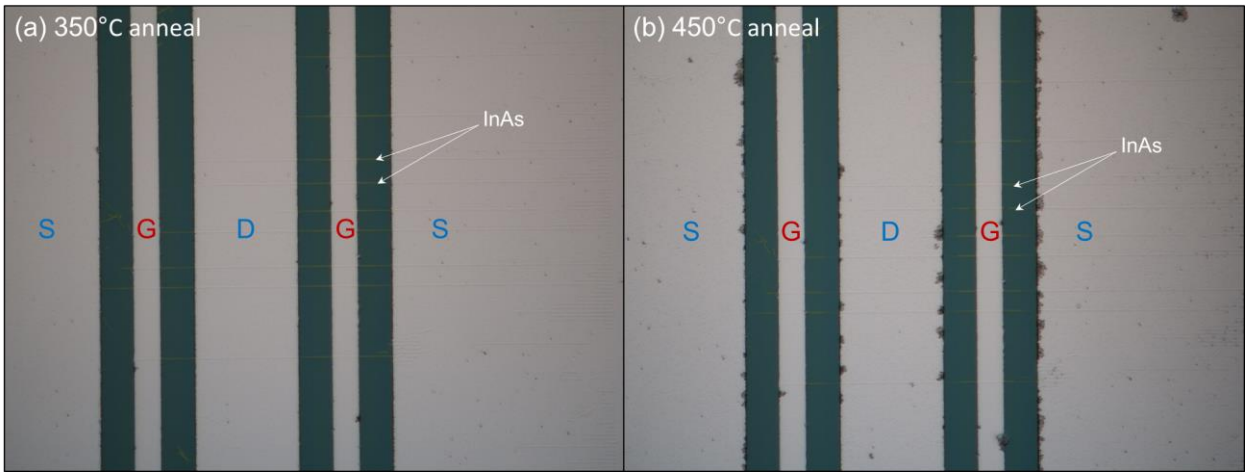


Fig. 70. Optical micrographs of a fully processed *n*-InAs JL-FETs on SiO_2 after annealing in N_2 for (a) 30 min at 350°C and (b) 60 min at 450°C. Discoloration of the InAs near the metal lines is visible after annealing at 450°C.

To fabricate JL-FETs out of those InAs NRs successfully transferred to SiO_2, we deposited 5 nm Al_2O_3 by ALD for the gate dielectric, evaporated 200 nm Al for the gate metal, and evaporated 20/200 nm Ti/Al for the S/D metal. Before S/D metal deposition, we first hard baked the resist for 5 min at 150°C and opened the contact windows by removing any Al_2O_3 and native oxide covering the S/D regions in 1:100 BOE/H_2O for 5 min and 1:10 HCl/H_2O for 30 sec. This recipe was found to be adequate to remove up to 15 nm Al_2O_3 on Si test pieces without destabilizing the NRs or the photoresist[12]. One of our devices is shown in Fig. 70(a) after a 350°C anneal for 30 min in a Carbolite oven.

Unfortunately, none of our fabricated InAs JL-FETs on SiO_2 showed any semblance of conductive behavior as the measured current levels were on the order of picoamps or less, just like the previous GaAs JL-FETs. Some of our two-terminal devices displayed conductive behavior, but at very low current levels (tens of nanoamps over ±20 V of applied voltage). Fig. 71 shows the I-V characteristics for those two-terminal devices (resistors) which had a measurable response. In

[12] At higher BOE/H_2O concentrations exceeding 1:10, the photoresist would lift off from the substrate indicating poor adhesion even with HMDS and an aggressive hard bake.

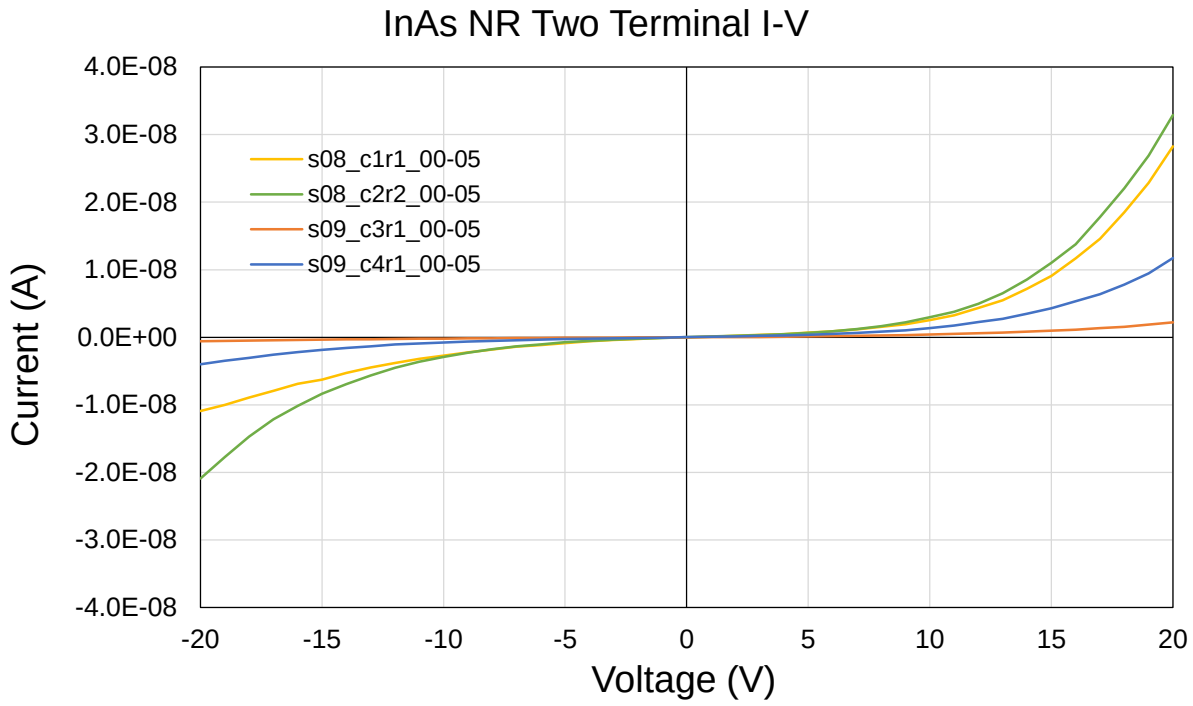


Fig. 71. Two-terminal I-V measurements performed on InAs NRs on SiO_2. The electrode separation is 5 μm and the InAs thickness is 15 nm for each device. The effective width of each device is unknown but lies somewhere between 0.5 and 2 μm.

each of the four devices, the resistor length is 5 μm while the effective resistor width is unknown[13] but lies somewhere between 0.5 – 2 μm. Based on the dimensions of the NRs, we expected currents on the order of 1–2 mA at a 1 V bias, whereas we measured less than 1 nA. The curves in Fig. 71 also resemble back-to-back Schottky diodes which is unexpected given the presumed ease of making ohmic contacts to *n*-InAs. It is possible that the Al_2O_3 covering the NRs was not completely cleared during the BOE etch which may have resulted in a residual tunneling barrier through the remaining oxide. Another possible explanation is that the NRs were damaged during pickup or transfer which may have compromised their physical or electrical integrity. We cannot determine if the NRs are, in fact, completely intact even in the regions which appear to be intact from visual

[13] The effective width depends on the number of InAs NRs that were successfully transferred within each array. Based on visual observation, it is difficult to assess how many of the NRs are actually intact after transfer, and it is even more difficult to determine whether or not each NR is electrically conductive. Since the resistors were designed for 4 NRs per device (each 0.5 μm wide), the effective width may vary anywhere between one to four NRs.

inspection. In other words, there could be microscopic breaks along the ultrathin NRs which are invisible under optical microscopy which lead to open circuits. A final possibility is that some PDMS residue has remained on the InAs after transfer which affected the electrical properties at the contact interface. Although we performed light oxygen plasma cleaning on the receiving SiO_2 substrate immediately after NR transfer, the plasma clean may not have been sufficient[14] to remove any remaining PDMS residue.

Although we were unable to demonstrate any electrically useful devices, the lessons from our experiments may be useful to other researchers working in the field of NTP-based HGI. We must stress that the lack of working devices here should not be construed as an indictment of NTP as a feasible or infeasible HGI process, especially since many other groups have successfully demonstrated working devices and heterogeneous circuits using transfer technology. Rather, what we have shown here is that despite the many challenges faced by this process, large arrays of extremely high aspect ratio NRs can be successfully transferred to foreign substrates which will be crucial to the success of NTP-based HGI as a commercially viable technology. In the next section, we show precisely why this is important as we introduce our HGI evaluation framework.

7.4 HGI Evaluation Framework

7.4.1 Objective of the Framework

Despite much ongoing research in developing HGI processes (including our own efforts from Section 7.3), there is little to no understanding of the overall impact feature-level HGI would have on near-future digital circuit generations. As a preliminary study, in Fig. 72 we show the

[14] A more aggressive piranha clean could not be used in this case because piranha etches GaAs and InAs.

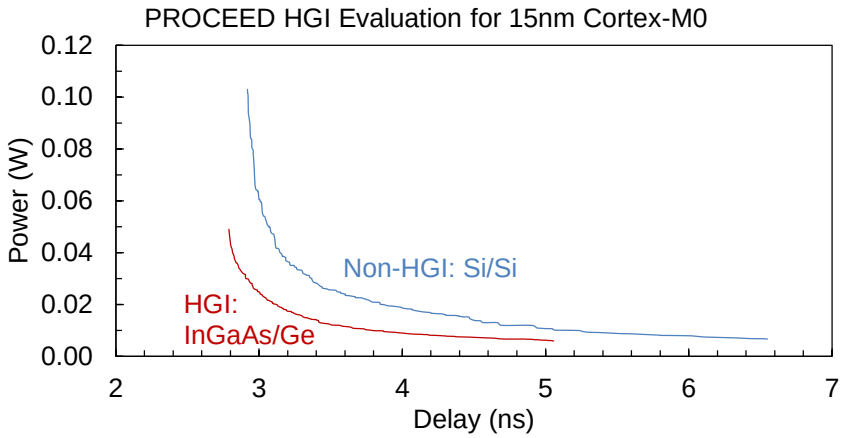


Fig. 72. Power-delay tradeoff for 15nm InGaAs/Ge and Si/Si built Cortex-M0 generated by PROCEED [154].

benefits of InGaAs- and Ge-based HGI IM-FinFETs over an all-Si design within a realistic processor architecture, as projected using the PROCEED evaluation framework [154][154]. Such assessments demonstrate that feature-level HGI could offer significant power savings at a given operating speed, although they do not consider layout area penalties that arise from HGI processes.

Here, we present for the first time a quantitative cross-layer study on the impact of NTP-based HGI versus Si-only technology on digital circuit performance and layout density. In the following sections we describe our HGI evaluation framework, considering achievable process capabilities (e.g., NTP overlay accuracy), intrinsic device performance, and circuit layout options. We will present the inverter- and block-level results of our cross-layer evaluation, using the specific case of VLSI circuits in 15nm IM-FinFET technology to compare the use of all-Si FinFETs with HGI of InGaAs and Ge as the NFET and PFET channel materials, respectively. We explicitly map the conditions in which this HGI technology holds an advantage over Si CMOS. The results of our simple and versatile framework should provide a tangible rationale for industry to seriously pursue HGI as a technology option in coming years.

To project the ultimate effects of HGI on future digital systems, we have developed a general methodology which we apply to the specific case of NTP-based integration. Our framework

is divided into three stages, 1) device simulation, 2) compact model calibration, and 3) circuit analysis, which are used to predict the benefits of an HGI (over non-HGI) implementation given the following data: 1) device specifications for a desired technology node, 2) an HGI process to implement the technology, and 3) representative circuit layouts for the technology, which will be used for benchmark comparisons. The following sections explain each part of the framework in more detail.

7.4.2 Device Modeling[15]

Because experimental data on scaled III-V MOSFETs is sparse, we use simulations to project *I–V* and *C–V* device performance at the 15nm node studied here. For maximal accuracy, we use the NEGF setup from Chapter 6 with some simplifications to perform quantum mechanical device calculations and capture important phenomena like ballistic transport and tunneling that cannot be fully modeled by conventional technology computer-aided design (TCAD). We use our in-house NEGF code to simulate 15nm Si and $In_{0.53}Ga_{0.47}As$ IM-FinFETs (not junctionless).

Our device structure is shown in the inset to Fig. 73, with physical gate length of 12.8 nm, channel thickness of 8.5 nm, oxide thickness of 0.68 nm, and supply voltage V_{DD} = 0.73 V. For all devices, gate work functions are adjusted to set the leakage current to 100 nA/μm. These values are taken from the ITRS projections for 15nm multigate devices. Our simulations assume ballistic transport, i.e., no scattering, which represents the upper bound of performance. Experiments show that devices are indeed approaching this limit as they scale, albeit more quickly for III-V compared to Si [73]. For *n*-type FinFETs, we perform effective mass simulations for Si and $In_{0.53}Ga_{0.47}As$ to extract device characteristics. We use three band k-p to simulate the Si PFETs; due to computational complications with the Ge band structure, we approximate Ge PFET devices by scaling the

[15] I am very grateful for Dr. Andrew Pan's valuable contributions to this section and for lending his NEGF simulations and modeling expertise to this work.

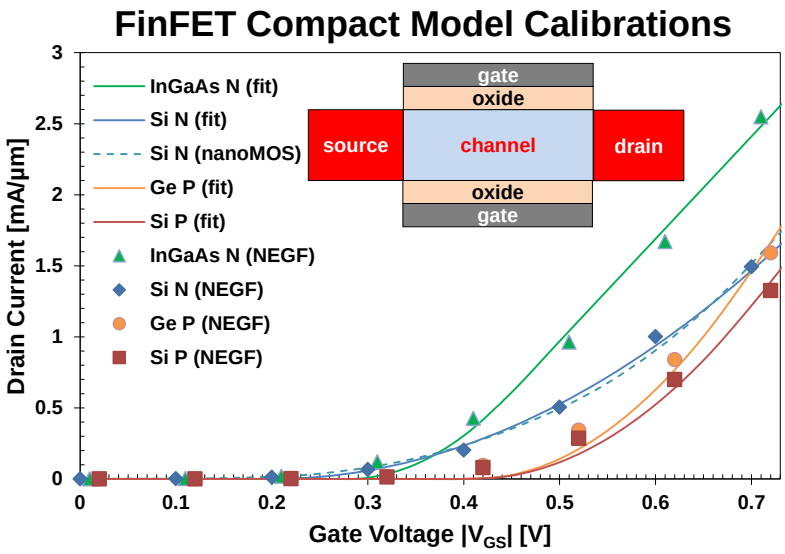


Fig. 73. NEGF (symbols) and model fit (lines) $|I_D|$-$|V_{GS}|$ curves for Si, Ge, and InGaAs double-gate FinFETs. The dashed line represents the NEGF Si NFET simulation using nanoMOS [155]. All simulations are with drain bias V_{DS} = 0.73 V. Inset: double-gate structure used for simulations.

Si characteristics by 20%, in accordance with other ballistic studies that show this enhancement ratio [156].

Lastly, we fit standard compact models to the simulated *I*–*V* curves for circuit delay calculations by adjusting parameters like mobility and saturation velocity. To validate our device simulations, we also compare our *n*-Si simulation with that performed using the standardized NEGF simulator nanoMOS [155] and observe close agreement. The characteristics and fits are shown in Fig. 73. We also extract the averaged off- and on-state capacitance for each device; for Si NFET and PFET and Ge PFET, this value is about 0.42 fF/μm, whereas it is about 0.27 fF/μm for InGaAs NFET. The reduced capacitance for III-V *n*-type devices is a well-known effect due to the conduction band DOS of such materials [157].

7.4.3 Alignment Error and Transfer Accuracy

The ability to overlay, align, and transfer heterogeneous features from one substrate to another with high accuracy will be critical to the success of any feature-level HGI transfer process. In Section 7.3.1 we showed that an overlay error $\sigma \approx 16$ μm was achieved in the experimental

demonstration of Fig. 62 using a standard contact aligner, and that much better tools would be needed to achieve the nanometer scale placement accuracies typically demanded in state-of-the-art VLSI technologies. Since there is no consensus on what σ values can be obtained (or will be needed) from NTP for use in future technologies, we surmise expected values of σ from 3 nm up to 50 nm for use in our framework, representing possible field-size "step and transfer" scenarios using state-of-the-art tools [143] derived from nanoimprint lithography (NIL). Recent NIL demonstrations [158] have shown minimum overlay errors of $3\sigma \cong 10$ nm for templates up to 2×3 cm^2 fields, so our projected σ values for NTP in this work should be reasonable—if not conservative—based on the similarities between NIL and NTP.

Both NIL and NTP are contact processes which use physical contact to either form or transfer patterned features to a substrate. In NIL, a mold containing the feature to be printed on the receiving substrate is physically pressed onto a UV-curable liquid resist layer on the substrate which results in displacement of the resist to conform to the mold's patterned shape. After the mold and resist are contacted, the resist is cured in light to solidify it and the mold is removed. Because NIL is a contact process, the overlay tolerance must be well controlled since features are printed at a 1:1 ratio with no magnification. The overlay accuracy depends on the precision of the stage movement, uniformity of the resist layer and the flatness of the mold and substrate surfaces [159], with the best demonstrations to date reaching $3\sigma \cong$ 10–15 nm [158]. Using a "step and imprint" technique [159], [160] the mold template can cover an entire field to balance throughput and accuracy over a large area.

In NTP, a soft adhesive stamp is used to pick up patterned structures from one substrate and transfer them to another. Unlike NIL, there is no actual lithography during the transfer process. Like NIL, however, the printing process relies on physical contact between two surfaces meaning

overlay accuracy will depend on flatness of the receiving substrate and the stamp containing patterned structures. In our experiments from Section 7.3, the PDMS stamp could vary in thickness by several hundred μm over a region of several cm^2. This can severely affect the alignment process due to limited depth of field in the equipment optics; when coupled with deformation of the stamp during contact transfer, the achievable overlay accuracy over large areas may be substantially limited compared to what is theoretically possible based on the (x, y, θ) precision of the stage movement. Because of this, it is reasonable to expect that NTP overlay accuracies based on our current experimentation capability may not yet reach those of the best NIL demonstrations to date. The reader should bear in mind, however, that engineering of the stamp properties may substantially reduce the severity of these issues, especially compared to what has/can be demonstrated in academic laboratories.

7.4.4 Transfer Yield and Performance Loss Considerations

Besides overlay accuracy, the transfer yield must be high enough to ensure that the process is reliable for commercial use. Since NTP as an HGI enabler is still in the early stages of research and development, reliable data about transfer yield is currently sparse. The authors in [134] claimed 87%, 95%, and 99% transfer yield[16] in their experiments for GaN, GaAs, and Si microribbons transferred to plastic substrates, indicating promise for this technology. The transfer yield from our GaAs on Si/SiO_2 demonstration in Fig. 66, however, was less impressive with an estimated transfer yield < 10%. As mentioned in Section 7.3.2, potential reasons for lackluster yield include microscopic variations in undercutting rates, bending stresses, poor adhesion strength between the stamp and semiconductor surface, and aspect ratio (AR) constraints resulting from the

[16] The concept of transfer yield is somewhat ambiguous since there is no universal definition which describes a "successful" transfer. Despite the high yields claimed in [134], close inspection of the images provided by the authors reveals a substantial number of geometrical defects in "successfully" transferred features.

limited structural integrity of nanoscale features during undercutting (and possibly suspension), pickup, and transfer. However, the length/width AR of our transferred GaAs NRs in Fig. 66 was ~533:1, which is among the highest reported values to date and, to our knowledge, the highest result for sub-1 μm wide features. A deeper investigation of the yield loss mechanisms and potential routes for improvement thereof are subjects of ongoing research, the results of which are expected to give more insight into what HGI circuit layout methodologies should be selected to enable more robust designs.

Even if 100% transfer yield can be achieved, the quality of transferred materials may be degraded after the stamping process. For example, the backside interface between the transferred fins and the receiving substrate could exhibit a higher density of interface traps due to poor bonding quality between the different materials, resulting in higher leakage current and parasitic capacitance. Since the amount of degradation will very likely be material- and process-dependent, it is difficult to quantify these effects without detailed experimental analysis. Some evidence suggests that NTP does not appreciably degrade the front-side interface between the channel and gate dielectric in terms of measured subthreshold characteristics from InAs-on-insulator FETs fabricated through a similar process [131], but more extensive studies will be needed to support this finding, especially regarding the backside interface properties. These topics remain the subject of ongoing research on our part.

7.4.5 HGI Impact on Circuit Layout and Design Rules

All HGI circuit designs face two new complications: a potential loss in intrinsic device performance (a likely problem for heteroepitaxy-based HGI due to crystal defects) and a reduction in layout density (particularly important for transfer-based HGI due to overlay accuracy limitations). The former effect can be accounted for by adjusting the device models presented in Section

7.4.2, but this is not easy to predict without extensive experimental data on HGI process-induced degradation of device characteristics. On the other hand, density loss can be easily accounted for by adjusting layout design rules, given some knowledge of the NTP overlay accuracy. Since we are mainly concerned with transfer-based HGI in this study, we will assume an ideal case where no loss in device performance occurs and focus on the layout area penalty from the NTP process. For the majority of this study, we assume that the NTP process occurs with 100% transfer yield; that is, no fins are missed or broken during the pickup and transfer steps. This is certainly optimistic, but it allows us to set an upper limit for the foreseeable gains from HGI. We do allow for misalignment of fins, however, resulting in "alignment yield" < 100%. We will not explicitly consider any rotational (θ) misalignment in this study, though its effects can be absorbed into additional x and y translational misalignments.

A simple FinFET inverter layout is shown in Fig. 74, where the PFET fins have been transferred with a one-sigma overlay error of $\pm\sigma$ and must satisfy three conditions: 1) the PFET fins must not land too close to the NFET fins, 2) all fins must be contacted by the drawn source, drain,

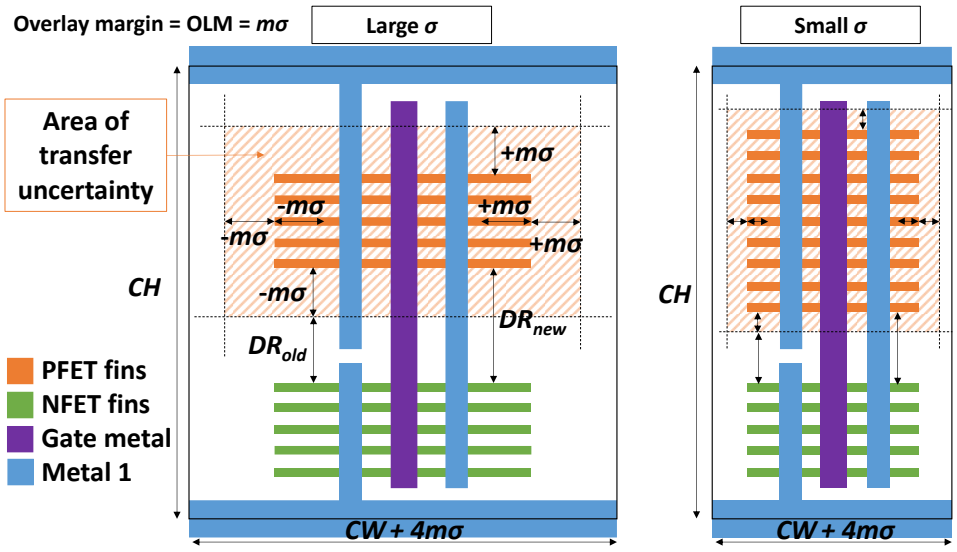


Fig. 74. Schematic layouts for heterogeneous FinFET inverters from NTP without fin trimming. The area of transfer uncertainty indicates the region where PFET fins can land due to misalignment.

and gate lines, and 3) all fins must lie within the cell boundaries. Each of these conditions imposes extra constraints on the appropriate design rules.

Separation of the PFET and NFET fins ensures that they do not overlap during or after transfer, causing device failure. In non-HGI layouts, a design rule setting the minimum distance DR_{old} between NFET and PFET fins exists due to the masked diffusion or implantation steps for the two devices; however, this minimum distance is not too large (~35 nm for the 15nm node) since it is set by lithography. In HGI layouts, however, the new minimum separation DR_{new} is increased by some multiple m of the transfer overlay accuracy σ, which may be significantly larger (~10 to 100+ nm). In other words, $DR_{new} = DR_{old} + m\sigma$. Determination of m is not straightforward and directly impacts the resulting alignment yield and area penalty at the cell level, as we will see later. Our approach for choosing m is detailed in Section 7.4.6.1. Conditions 2 and 3 impact the HGI layout area penalty differently depending on the presence of a trim step after fin transfer (Step 8 in Fig. 60), meriting a separate discussion.

7.4.5.1 HGI without Fin Trimming

The requirement that all fins be properly contacted has two consequences. First, the fin length must be extended by $m\sigma$ on each end, meaning the minimum fin length increases by $2m\sigma$ in order to guarantee proper electrical contact when a $\pm m\sigma$ horizontal HGI misalignment occurs. This also means the minimum cell width (CW) must increase by $2m\sigma$ to accommodate the longer fins when HGI is used. Second, to absorb any vertical HGI misalignment, the maximum number of transferable PFET fins per cell is reduced to a value dictated by the fin pitch, the minimum fin-to-metal 1 (M1) overhang, and the minimum M1-to-M1 separation. The end result is that fewer PFET fins can be transferred within a minimum size cell when HGI overlay accuracy is poor compared

to the non-HGI case; a “stronger” PFET will require more transistor folding and consume a larger cell area.

Finally, to enforce the cell boundaries and account for any vertical misalignment, the minimum distance from the PFET fins to the top of the cell becomes $m\sigma$. This sets another limit on the number of PFET fins that can be transferred within a minimum sized cell. More catastrophically, the cell boundary condition also forces the cell width to increase by an additional $m\sigma$ on each side for a net increase of $2m\sigma$. Adding this to the $2m\sigma$ penalty from using longer fins means the width of *every* cell must increase by a total of $4m\sigma$, absent fin trimming. For instance, if σ = 50 nm and m = 2 (for 95% alignment yield), every cell would widen by 400 nm, thereby increasing cell area by more than 5× over a 15nm non-HGI design.

7.4.5.2 HGI with Fin Trimming

Fin trimming (see Step 8 in Fig. 60) effectively removes the impact of lateral misalignment on layout except for the cells at the ends of a row. This is because lateral misalignment will only appear at the left and right fin ends as shown in Fig. 75, which will inevitably be removed after the trim. Within each row there is no need for the fins to be longer than normal to guarantee electrical contact, nor is there a need for extra room in the $\pm x$ direction to keep neighboring transistors isolated since the trim step guarantees it. Thus, the cell width does not increase (discounting transistor folding) to accommodate HGI overlay.

The only area penalty incurred is the addition of two dedicated empty regions (at least $2m\sigma$ in length) which absorb the misalignment penalty at the very ends of each transferred fin. Since the empty regions can sandwich many active cells within a row, this area penalty is amortized across the cells, mitigating the per cell penalty, and reduces as the transferred fin length increases.

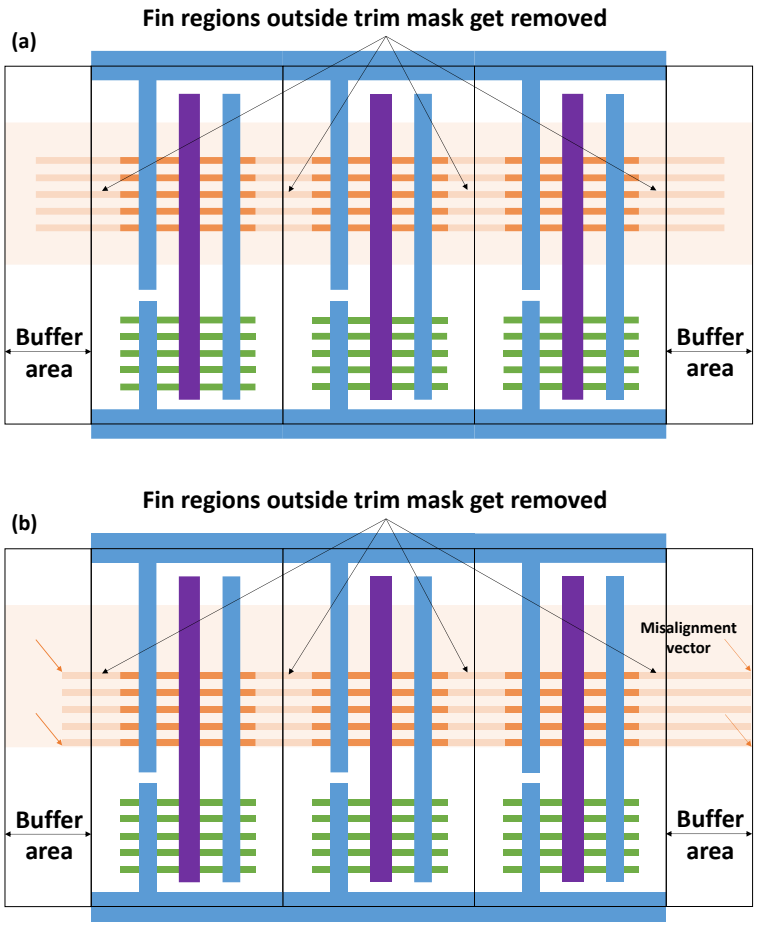


Fig. 75. (a) Schematic layout for a row of heterogeneous FinFET inverters made with NTP and fin trimming. (b) The effect of transfer misalignment with fin trimming is now absent within each cell except at the buffer areas on ends of a row.

Most likely, however, arbitrarily long fins cannot be transferred with good yield due to complications from microscopically variable undercutting rates before fin pick-up, peeling forces during transfer, and stamp surface topography (recall the process challenges faced in Section 7.3). We speculate that transfer yield may be correlated to the fin length/width AR, limiting the transferrable fin length and per-cell penalty reduction. Unfortunately, exact constraints on the AR are not clear at this point due to limited experimental evidence; this will be revisited later in Section 7.4.6.3.

Ultimately, compared to non-HGI circuits of equal performance, circuits using HGI will incur a layout density hit that is dependent on σ as well as the number of fins in each cell (i.e., the cell strength). As an example, for a given cell height, a minimum size inverter with just one NFET

and PFET fin can tolerate a larger misalignment due to the large amount of empty space in the cell, whereas a cell containing more NFET and PFET fins can only tolerate a small misalignment before design rule violations occur. Consequently, for a given HGI process (i.e., a given value of σ), only some circuit cells will incur a layout area increase.

Finally, the reader may note that all area penalties mentioned originate only from the PFET transfer. The reason is that the NFET fins are transferred to the receiving wafer before any other patterns are formed, so they serve as the reference to which all other features (PFET fins, gate/interconnect lines, etc.) are aligned. As such, the alignment-related penalties discussed here are assumed to only apply to PFETs.

7.4.5.3 Circuit Level Evaluation[17]

In our framework, we use UCLA Design Rule Evaluator (DRE), a free online tool [161], to generate 15nm FinFET circuit layouts using modified design rules to account for the HGI-related penalties. For simplicity, the 15nm design rules are first obtained from a scaled version of an existing 45nm [45] planar process where all dimensional quantities are scaled by 15/45 = 33%. Once a nominal set of rules is obtained, a subset is modified to account for the different methods of FET formation: physical transfer in the HGI process, and standard lithography plus etch for the non-HGI process. The actual rule values used in our study will be discussed later in Section 7.4.6.1.

With the design rules in place, we synthesize a 15nm cell library using Nangate Open Cell Library [45] as a template[18] and scale all transistor sizes to match the 15nm node. All FinFETs have gate length L_g = 13 nm and effective width $2N \times H_{fin}$, where N is the number of fins per

[17] We sincerely thank the UCLA NanoCAD Laboratory for providing assistance with the DRE program and especially Shaodi Wang for performing the chip-level studies described in the next section.

[18] While the 15nm library used is not derived from an actual commercial FinFET library (bearing in mind that no such library has been made publicly available), we believe our findings should still be useful to the design community even if the reported results are based on projected inputs.

transistor and H_{fin} = 17 nm is the fin height. After the cell layouts are generated, switching delays are estimated for each cell using the fitted compact model parameters discussed in Section 7.4.2. Using this simple model, we can rapidly compare the cell-dependent impact of different HGI technology and design rule scenarios without brute force circuit simulations over an entire library. Once the cell library is characterized for each type of process (HGI and non-HGI), we compare the relative delay–area and delay–power impact across a few benchmark designs for a full chip-level HGI evaluation.

7.4.6 Projected HGI Benefits

7.4.6.1 Setting the HGI Design Rules

In Section 7.4.4, we noted that DR_{new} must exceed DR_{old} by at least $m\sigma$ to ensure good alignment yield (AY) without consuming excessive area. Assuming a Gaussian distribution for the overlay error, we have alignment yield AY(OLM) = erf($m/\sqrt{2}$) where erf is the error function and OLM ("overlay margin") = $m\sigma$. This is plotted in Fig. 76(a) for σ = 3, 25, and 50 nm, representing expected NTP capabilities as discussed in Section 7.4.3. As shown in Fig. 74, OLM essentially represents the extra space needed on all sides of the PFET fins to account for HGI misalignment. Note that an increase in OLM only results in larger cell area if transistor folding becomes necessary for a given cell height (CH) and cell strength (number of fins). We determine OLM by optimizing the alignment yield per unit average cell area for a given HGI process (i.e., value of σ); this is analogous to optimizing the design rules to obtain the maximum number of good dice per wafer. Since different cell types have different optimum OLM, we consider a reduced size MIPS processor as our benchmark and compute an average cell area weighted by the number of cell instances of each type. Based on Fig. 76(b) and (c), for σ = 3 nm, 25 nm, and 50 nm, the optimum values of

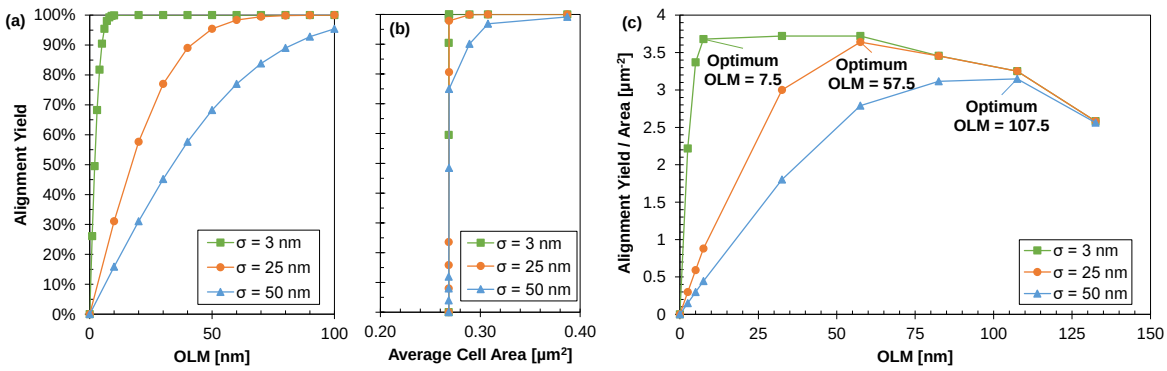


Fig. 76. (a) Probability of successful fin placement as a function of transfer misalignment and allotted overlay margin. (b) Alignment yield versus average cell area in reduced MIPS processor. (c) Optimal OLM value search to maximize alignment yield per cell area.

Table 23. Modified 15nm Design Rules for Different Process Scenarios

Process	P-N spacing (intra-cell) [nm]	P-P spacing (inter-cell) [nm]	Minimum cell dimensions [nm]
Non-HGI	H: n/a	H: 72	H: 72
	V: 35	V: 72	V: 506
HGI (no trim)	H: n/a	H: 72+2OLM=187	H: 72+4OLM=302
	V: 35+OLM=92.5	V: 72+2OLM=187	V: 506
HGI (trim)	H: n/a	H: 72	H: 72
	V: 35+OLM=92.5	V: 72+OLM=129.5	V: 506

Note: For HGI processes, σ = 25 nm and OLM = 57.5 nm are used. "H/V" specifies design rule value in horizontal/vertical direction.

OLM are 7.5 nm (m = 2.5), 57.5 nm (m = 2.3), and 107.5 nm (m = 2.15) respectively, corresponding to alignment yield of 97, 98, and 99%. As a rule of thumb, it appears OLM ≅ 2σ is a good choice for the allotted overlay margin due to misalignment.

Table 23 summarizes the modified design rules for HGI circuits assuming a transfer accuracy of σ = 25 nm as well as the baseline design rules for non-HGI circuits. Here, we have DR_{old} = 35 nm (introduced in Section 7.4.5) and OLM = 57.5 nm. The P-N spacing is the same as DR_{new} and accounts for misalignment in the –y direction, while the P-P spacing accounts for misalignment in the ±x and +y directions.

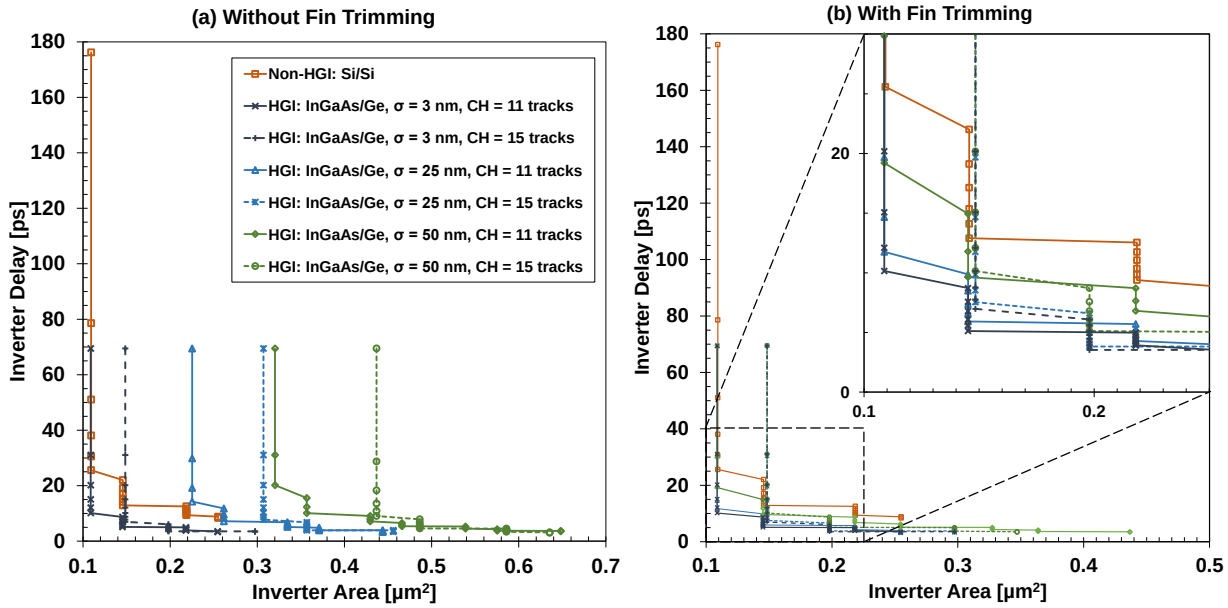


Fig. 77. Delay versus area for 15nm InGaAs/Ge (HGI) and Si/Si (non-HGI) inverters for different σ and CHs (a) without fin trimming and (b) with fin trimming. The inset is a magnified view of the dashed region in (b).

7.4.6.2 Inverter Delay vs. Area Evaluation

With the framework so far in place, we examine the tradeoffs between delay and area for FinFET inverters of varying strength (i.e., number of fins) implemented in either InGaAs/Ge (HGI) or Si/Si (non-HGI) processes: the notation "A/B" refers to a cell using material "A" for the NFET and "B" for the PFET. For each set of design rules per process scenario, we obtain a series of inverter delay—area curves such as those shown in Fig. 77. Starting from the top of each curve and moving downward, each successive marker represents an increment in the number of PFET and NFET fins in the inverter, beginning with 1 and ending at 20 fins, mapping out the inverter's delay and area as a function of cell strength from 1X to 20X. The cell height (CH) in each case is either 11 or 15 (Metal 3) tracks for InGaAs/Ge inverters, while for Si/Si inverters CH is fixed at 11 tracks.

When fin trimming is neglected, there is no point at which any of the $\sigma \geq 25$ nm HGI configurations holds a clear advantage over the Si/Si baseline, as evidenced by comparing the curves in Fig. 77(a) at a given delay value: the baseline can always provide the same delay while

consuming a smaller footprint. We also observe that taller cells pay a larger initial area overhead compared to shorter cells but become more attractive as the cell strength increases, since fewer transistor folds are needed in a taller cell. High performance circuit blocks using many fins per transistor can be designed with taller cells to minimize folding, while low power blocks using only a few fins per transistor can be designed with shorter cells to minimize the area overhead. As expected, large σ values result in larger areas due to more frequent folding. When σ = 3 nm, however, most of the InGaAs/Ge curve lies well below the Si/Si baseline, indicating substantial benefits. This represents the upper limit of what HGI can offer assuming such accurate transfers are possible.

When fin trimming is included, the layout area penalty is reduced such that all the HGI delay–area curves in Fig. 77(b) have at least some advantageous regions that lie beneath the baseline. In fact, for σ = 25 nm and CH = 11 tracks, nearly the entire curve lies below the Si/Si baseline with the InGaAs/Ge inverter able to provide >50% reduction in delay for the same area. For σ = 50 nm and CH = 11 tracks, InGaAs/Ge still offers benefits, but the constant-area delay reduction is only roughly 25%. For taller cells (CH = 15 tracks), the initial overhead represents an extra 35% area cost for the weakest cells but starts to pay off once the cell strength exceeds 10X when σ = 25 nm and 6X when σ = 50 nm. The benefits of migrating to a taller CH (11→15 tracks) are more apparent when σ is larger: folding frequency is reduced from every six to every eight fins (25% less folding) when σ = 25 nm, but from every three to every six fins (50% less folding) when σ = 50 nm. Depending on the balance of weak and strong cells in the circuit design, it may be advantageous to design with taller cell heights everywhere when adopting HGI, especially if the transfer accuracy is poor.

Table 24. Area, Delay, and Power of HGI Standard Cells Compared to Non-HGI Cells.

A) Summary of area comparison

Cells in library with area overhead from HGI	Average weighted area overhead (MIPS)
24 of 114	6.6%

B) Summary of delay comparison

Cells in library with delay reduction from HGI	Average weighted delay reduction (MIPS)
114 of 114	62%

C) Summary of power comparison

Cells in library with power reduction from HGI	Average weighted power reduction (MIPS)
114 of 114	18%

D) Standard cells with area overhead

Cell	Area overhead	Cell	Area overhead
AOI222_X4	94.6%	OAI222_X2	21.9%
NOR4_X2	85.5%	BUF_X4	19.7%
INV_X8	74.7%	AND2_X4	16.4%
NAND2_X4	74.7%	OR2_X4	16.4%
AOI222_X2	49.3%	OAI222_X4	15.3%
NOR4_X4	46.4%	AND3_X4	12.3%
INV_X16	40.3%	OAI21_X4	12.3%
NAND3_X4	36.4%	OR3_X4	12.3%
BUF_X8	33.1%	AND4_X4	9.9%
BUF_X16	33.0%	OR4_X4	9.9%
INV_X4	32.9%	INV_X32	8.2%
NAND4_X4	31.4%	BUF_X32	6.9%

7.4.6.3 Block Level Evaluation

For circuit block analysis, we again investigate designs involving either non-HGI (Si/Si) or HGI (InGaAs/Ge) configurations. We modify the digital circuit backend flow to properly account for the area adjustments induced by NTP. For misalignment, following the arguments above, we have seen that the use of fin trimming essentially eliminates OLM in the *x* direction and any cell area penalties arise only from *y* direction OLM and added transistor folding. We use UCLA DRE [161] to generate all HGI standard cells based on the Nangate Open Cell Library templates [45] with calibrated design rules including misalignment penalties as discussed in Sections 7.4.5.3 and 7.4.6.1. In all results to follow, the use of post-transfer fin trimming is assumed. The area overheads, delay reduction, and power reduction of HGI cells compared to non-HGI are given in Table 24, assuming OLM = 50 nm (roughly corresponding to σ = 25 nm) and CH = 11 tracks. We

see 24 out of the 114 HGI cells incur an area penalty (of up to 94.6%) and an overall 6.6% area increase is seen after weighted averaging based on usage in MIPS. The stronger drive currents and lower capacitance from InGaAs/Ge HGI result in lower delay and power for all 114 standard cells in the library. Circuit benchmarks are then synthesized in a commercial synthesis tool using these standard cells.

In addition to the OLM requirements, there may be limits to the transferable fin length imposed by the fin aspect ratio as mentioned in Section 7.4.5.2. In practice, this means that sets of long but finite fins will be transferred, with additional gaps between adjacent fins. However, these sets are transferred simultaneously without incurring relative OLM, so that the only added area overhead comes from gaps between the sets. To include these gaps, prior to cell placement we insert a grid of dummy filling cells on the placement rows separated by a distance equal to the maximum allowed fin length[19] (MAFL) as shown in Fig. 78. These filling cells are temporarily fixed in the layout and the design cells are then placed using a commercial placement tool. The filling cells guarantee that the fin length, which is the width of the connecting cells, does not exceed

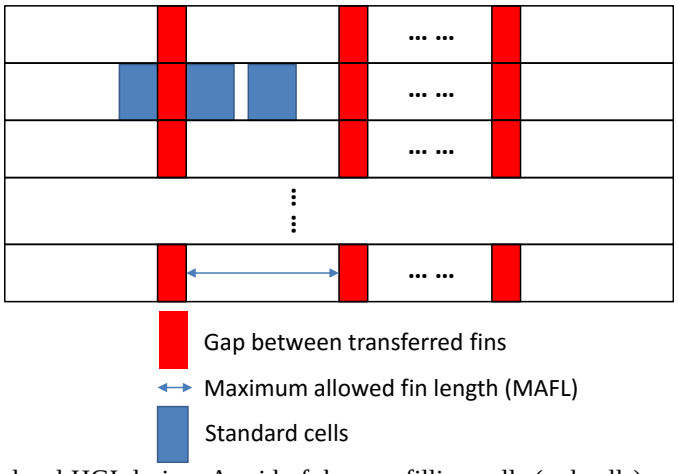


Fig. 78. Protocol for block-level HGI design. A grid of dummy filling cells (red cells) are inserted pre-placement to represent the effect of finite fin length, and standard cells (blue cells) are then placed in between the filling cells.

[19] The MAFL represents the hypothetically longest fin length which can be transferred with 100% yield (which is assumed throughout this work), considering the process challenges mentioned in Section 7.4.3. The higher the MAFL, the better it is.

the MAFL. The width of the fill cells is set to the minimum gap required in the transfer process (2OLM). After placement, the dummy cells are removed and routing is performed.

In the block-level simulations, multiple delay constraints are set during circuit synthesis using different technology libraries. Synthesized circuits are then placed and routed (P&R) within a fixed-size die with a grid of filling cells. This die size accommodates the Si/Si baseline design with 80% utilization. We first compare the *pre*-P&R delay versus area tradeoffs of HGI and non-HGI implementations in MIPS and AES benchmark designs to estimate the impact of misalignment-induced penalties. Then the *post*-P&R delay and power are compared for the same benchmarks, which also includes the penalty from reserved areas (gaps) between adjacent but disconnected sets of transferred fins. The results to follow assume that MAFL = 5 μm and the gap between adjacent sets of fins is given by 2OLM. We consider two situations for HGI: 1) an ideal scenario of σ = 3 nm (OLM = 6 nm) which essentially means no penalty[20] from misalignment, and 2) a more realistic scenario of σ = 25 nm (OLM = 50 nm). In both cases the alignment yield is 95% according to the analysis in Section 7.4.6.1. By comparing these two cases with the non-HGI scenario, we can separate the gains in chip performance/density due to the use of InGaAs and Ge as channel materials from the degradation due to the transfer technology.

[20] The MAFL is assumed to be infinite for σ = 3 nm since a 6 nm OLM overhead is already easily satisfied by the default (non-HGI) standard cell design rules, and thus no extra "filling cells" are ever needed in Fig. 78 nor do any of the HGI standard cells require enlargement from their default non-HGI sizes.

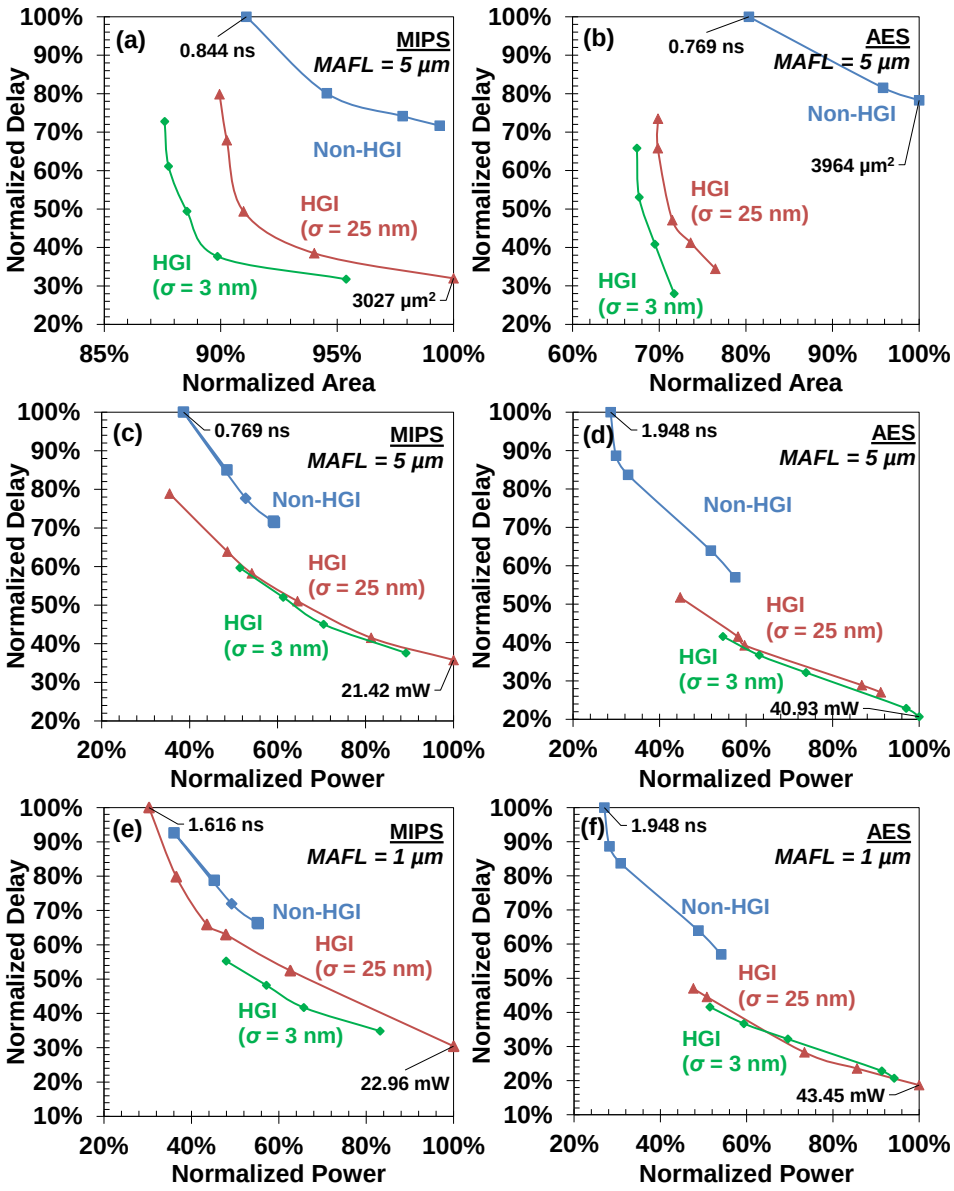


Fig. 79. Post-synthesis (pre-P&R) normalized delay and area of (a) MIPS and (b) AES designs. Post-P&R normalized delay and power of MIPS and AES designs with MAFL of (c,d) 5 μm and (e,f) 1 μm, respectively. In each panel the reported data is normalized to the largest observed delay, power, or area values as indicated by the data labels. The design rules (i.e., OLM values) are chosen to ensure 95% yield in all cases.

In Fig. 79(a)–(b) we present the *pre*-P&R normalized delay–area curves for MIPS and AES benchmarks under the HGI and non-HGI scenarios introduced earlier. Clearly, the InGaAs/Ge design outperforms the non-HGI design in both delay and area efficiency. From Section 7.4.2, InGaAs and Ge both offer stronger driving current than Si, while InGaAs also possesses lower intrinsic capacitance than Si due to its lower density of states. These advantages outweigh the

higher area overheads (i.e., from OLMs) due to transfer misalignment since weaker (smaller) and/or fewer cells can be used in the design while still meeting the performance target. For instance, to achieve the same target clock period of 600 ns in AES design synthesis, InGaAs/Ge (σ = 25 nm) requires only 1358 buffers and inverters, while the non-HGI design needs 2845 buffers and inverters. This is exemplified in Fig. 79(b), where the HGI designs can actually show chip area savings compared to the non-HGI case despite the higher penalties from transfer misalignment. The benefits of InGaAs and Ge are even more apparent for the ideal σ = 3 nm scenario, which represents the full potential of HGI technology.

In Fig. 79(c)–(f), the *post*-P&R delay and power tradeoffs are compared for the same benchmark designs with MAFL = 5 μm and 1 μm. The penalty arising from the gaps between adjacent sets of transferred fins generally leads to higher interconnect delay and power. Again, the intrinsic performance advantage from using InGaAs/Ge-based HGI overwhelms the overhead area penalties, leading to much better performance and power efficiency compared to the non-HGI design even for short MAFL. We note that in (c), (d), and (f), the σ = 3 and 25 nm HGI designs give very similar performance within a fixed-size die which suggests that the extra penalties to routing from the dummy filling cells in Fig. 78 are insignificant.

Finally, we explore the HGI design impact resulting from constraints on the maximum allowed fin length due to NTP challenges. We place designs synthesized with the same delay constraint in a fixed-size die with MAFL ranging from 1 to 20 μm, representing fin ARs of 120:1 to 2300:1 for 15nm FinFETs. For comparison, the experimentally demonstrated AR in Fig. 66 is 533:1. In Fig. 80, the total wire length decreases when longer fins can be successfully transferred. Additionally, the wire length drops quickly with incremental improvement in fin length for short fins, but then saturates for longer fins. The reduction in total wire length with longer MAFL is

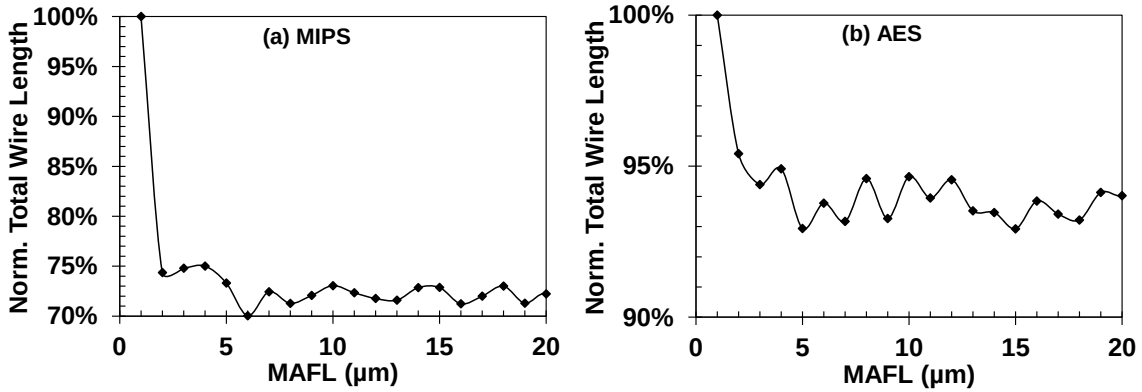


Fig. 80. Total interconnect length as a function of maximum allowed fin length for HGI-based (a) MIPS and (b) AES designs.

more apparent in MIPS compared to AES; this is because fewer cells in MIPS are connected by long metal lines, unlike AES. This also explains why the σ = 3 nm HGI designs showed more improvement compared to the σ = 25 nm designs in Fig. 79(e) but not in (c), (d), or (f): the short MAFL of 1 μm leads to a routing bottleneck in MIPS when there is significant transfer misalignment (i.e., σ = 25 nm) and hence leads to tempered performance gains. This illustrates how the transfer capabilities can have a stronger impact on designs which normally suffer from higher routing congestion. A maximum allowed fin length of 5 μm or more (AR > 600:1) should not pose a bottleneck for HGI except for the densest designs.

For the materials and models considered, full InGaAs/Ge HGI shows the best characteristics, though naturally other material and design scenarios remain. While full HGI offers the most benefits in terms of performance, power, and area over non-HGI, the higher cost of implementing a two-step transfer process may pose a legitimate manufacturing concern. Our constant-leakage, constant-voltage results also do not consider the possibility of using HGI to scale supply voltage, which opens up more possibilities for performance optimization. While the quantitative results will change somewhat depending on chip architecture and utilization ratio, these results clearly illustrate the attractiveness of NTP-based HGI for near-future digital designs and provide motivation for the development of more sophisticated HGI design methods and models.

7.4.7 Evaluation Summary

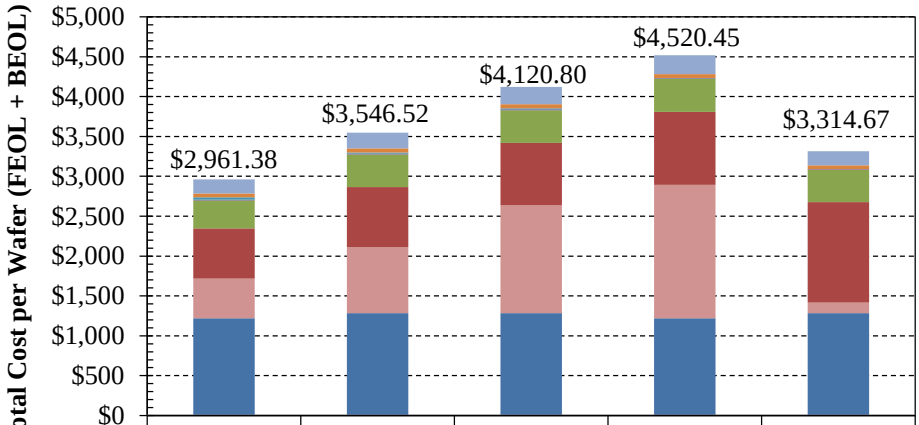
Our evaluation framework reveals that substantial improvements in circuit delay and power can be obtained using heterogeneous designs while trading off layout area. HGI cell area grows in response to more stringent design rules stemming from nanotransfer overlay misalignment, resulting in more frequent transistor folding and larger minimum cell widths. Fin trimming significantly reduces the lateral misalignment penalty and will likely be mandatory for HGI adoption. Designing strong cells with taller cell heights to reduce the folding frequency can also be beneficial when σ is large, despite the higher initial area overhead. Using InGaAs and Ge as heterogeneous materials to replace Si, sizeable reductions in processor delay (up to 40%-50%) and power (up to 15%-20%) are observed in HGI-based designs. Despite additional area overheads stemming from transfer misalignment, HGI designs actually consume less overall area compared to their non-HGI counterparts because some cells now require fewer fins than before to provide the same cell strength and designs will require fewer buffers to minimize critical path delays. Although we considered HGI of IM-FinFETs in this case study, the conclusions drawn should also apply to JL-FETs as well. Our findings provide strong motivation for the process and design communities to pursue feature-level heterogeneous integration as a viable option for nanoscale semiconductor fabrication.

7.5 Cost Analysis

Besides evaluating the performance gains from HGI adoption, it is also worth examining whether or not there may be a substantial cost overhead in transitioning from a standard non-HGI process to an HGI one. For example, compared to a standard Si FinFET process, will a commercial transfer-based InGaAs/Ge HGI process incur a higher cost per wafer? Obviously, we cannot answer this question with any true certainty due to immaturity of HGI technology and the usually

confidential nature of real-world cost-of-ownership (COO) information. However, we can make use of simple COO estimates based on the model developed by Wen and Chui [162] which was originally formulated to compare the manufacturing costs of 22nm FinFET and JL-FET technologies from both SOI and bulk Si wafers. With proper modifications to the front end of line (FEOL) process sequence to account for the different material and toolset requirements needed for HGI technology, we can adapt the same model to perform a cost comparison between non-HGI and HGI FinFET technologies. Due to complexity of the cost model, we omit a detailed explanation of the framework and kindly refer the reader to the original manuscript by Wen and Chui for a more complete description.

A comparison of the estimated cost breakdowns for implementing 22nm JL-FET technology in non-HGI and HGI process scenarios is shown in Fig. 81, while a brief summary of the



	Non-HGI	NTP - Si/Ge	NTP - InGaAs/Si	NTP - InGaAs/Ge	NHE - InGaAs/Ge
Other	$177.47	$197.47	$217.47	$236.96	$176.96
CMP	$49.03	$49.03	$49.03	$49.03	$49.03
Doping	$26.95	$14.15	$14.15	$0.00	$0.00
Anneal	$12.20	$12.20	$12.20	$6.96	$6.96
Etching	$348.89	$408.08	$408.08	$420.79	$403.97
Deposition	$628.70	$754.77	$784.06	$913.57	$1,261.93
Substrate	$500.00	$825.00	$1,350.00	$1,675.00	$130.00
Lithography	$1,218.13	$1,285.81	$1,285.81	$1,218.13	$1,285.81

Fig. 81. Estimated cost breakdown to implement 22nm FinFET technology in different process scenarios. For HGI processes, the integrated material pair is realized by either nanotransfer printing (NTP) or nanoheteroepitaxy (NHE).

Table 25. Condensed Process Sequences for HGI and Non-HGI Options

Non-HGI	NTP - Si/Ge	NTP - InGaAs/Si	NTP - InGaAs/Ge	NHE - InGaAs/Ge
12" SOI	12" SOI, 6" Ge	12" SOI, 4" InP	4" InP, 6" Ge, 12" Si	12" Si
n-Si channel doping (ion implant) (*)	n-Si channel doping (ion implant) (*)	p-Si channel doping (ion implant) (*)	Si oxidation	Si oxidation
p-Si channel doping (ion implant) (*)	n-Si fin patterning (*)	p-Si fin patterning (*)	**Source substrate layer growth (p-Ge/AlAs and n-InGaAs/InAlAs)**	Open trenches (*)
n/p fin patterning (*)	n-Si dummy fin removal (*)	p-Si dummy fin removal (*)	**n-InGaAs fin pattern-ing** (*)	Grow p-Ge fins (*)
n/p dummy fin removal (*)	**Source substrate layer growth (p-Ge/AlAs)**	**Source substrate layer growth (n-In-GaAs/InAlAs)**	**n-InGaAs fin transfer to Si wafer**	Grow n-InGaAs fins (*)
	p-Ge fin patterning (*)	**n-InGaAs fin pattern-ing** (*)	**p-Ge fin patterning** (*)	Remove dummy p fins (*)
	p-Ge fin transfer to Si wafer	**n-InGaAs fin transfer to Si wafer**	**p-Ge fin transfer to Si wafer**	Remove dummy n fins (*)
	p-Ge fin trim (*)	n-InGaAs fin trim (*)	n-InGaAs fin trim (*)	
			p-Ge fin trim (*)	
HK+MG stack (*) x 2	HK+MG stack (*) x 2	HK+MG stack (*) x 2	HK+MG stack (*) x 2	HK+MG stack (*) x 2
Gate spacer (*)	Gate spacer (*)	Gate spacer (*)	Gate spacer (*)	Gate spacer (*)
Raised S/D (*) x 2	Raised S/D (*) x 2	Raised S/D (*) x 2	Raised S/D (*) x 2	Raised S/D (*) x 2
Contact formation (*)	Contact formation (*)	Contact formation (*)	Contact formation (*)	Contact formation (*)
BEOL M1-M9 (*) x 19	BEOL M1-M9 (*) x 19	BEOL M1-M9 (*) x 19	BEOL M1-M9 (*) x 19	BEOL M1-M9 (*) x 19

(*) denotes a lithography step is needed for the particular step.
Steps in **bold** are performed on the source substrate (wherever appropriate).

FEOL process sequences for each scenario is provided in Table 25. For the three NTP-HGI scenarios we consider the following possibilities: Si/Ge, InGaAs/Si, and InGaAs/Ge. For additional comparison, a fourth HGI scenario is also considered: InGaAs/Ge directly grown on Si by nano-heteroepitaxy (NHE) in patterned windows. In Table 25, we assume that once the NFET and PFET fins are successfully formed on the host substrate, the remaining process sequence (starting from the high-k dielectric and metal gate stack formation) will be more or less generic/common[21] to all of the scenarios under consideration.

From Fig. 81 we immediately see that NTP-HGI is expected to be more expensive than non-HGI for several reasons. The first and most obvious is the higher cost[22] of Ge and InP (for $In_{0.53}Ga_{0.47}As$) source substrates compared to Si. Second, the channel materials must be epitaxially

[21] Processing dissimilar channel materials on the host substrate would necessitate uniquely tuned recipes for each scenario, but for simplicity we assume that no *radical* changes to the process sequence will be needed, such as the introduction of extra lithography, deposition, or etch steps and/or other nonstandard procedures.

[22] Per wafer estimates obtained from different vendors: 12" Si = $130, 12" SOI = $500, 4" InP = $850, 6" Ge = $325.

grown on the source substrates by MBE or MOCVD which can be slow and expensive. Third, wafer sizes for exotic materials are usually limited to 6" or less in diameter, meaning the transfer process must be repeated multiple times (i.e., "scale up") to fully populate each 12" Si host wafer. Assuming the use of superlattices structures on the source substrates[23], this multiplies the total epitaxy cost by the number of periods required on each source substrate which is a function of the die per wafer (DPW) ratio between the different sized substrates. These three factors directly contribute to the greater expenses listed under the "Deposition" and "Substrate" categories in Fig. 81. The percentage cost increases for the NTP-HGI scenarios compared to non-HGI are: +20% for Si/Ge, +39% for InGaAs/Si, and +53% for InGaAs/Ge.

Interestingly, the cost increase of InGaAs/Ge NHE-HGI is relatively small at only +12%. The higher cost of NTP compared to NHE mainly arises from the exotic substrates that are needed; the only substrate required in NHE is a bare Si wafer. Although epitaxy is required in both NTP and NHE, in NTP the 6" Ge and 4" InP source substrates have less area compared to a full 12" Si wafer so they require fewer expensive consumables (in terms of precursor volume) for growth which is a major component of the total epitaxy cost [162]. However, the substrate cost savings greatly overwhelm the higher deposition costs and ultimately result in a smaller cost overhead compared to NTP-HGI. While this is positive news for NHE as a potential option for commercial HGI implementation, issues over growth quality due to lattice constant mismatch and low throughput from the MBE/MOCVD tools will still remain.

Before concluding, we remind the reader that many implicit assumptions about HGI-related tools and materials pricing were taken, but not explicitly stated[24] or justified, to obtain the

[23] See Fig. 61 and the corresponding discussion in Section 7.2.

[24] The reader is welcome to contact me directly for specific details about the cost figures and assumptions used in the model.

cost estimates presented in Fig. 81. For example, it is impossible to predict how expensive the automated alignment and transfer systems would be for NTP-HGI, and the prices of some required consumables (e.g., trimethylindium precursor for InGaAs growth) were unavailable to us; in such cases, we were forced to make conservative guesses simply based on perception of operational complexity or material rarity. The anticipated process throughput and yield from HGI are also uncertain and would affect the cost comparison as well. Given the myriad of unknowns, such real-world cost projections are (understandably) vague at best so the reader must take the presented results with caution.

7.6 Summary

We have shown that NTP can open a pathway to enable feature-level HGI in VLSI circuits. As an HGI process, NTP allows us to cointegrate dissimilar materials unfettered by the lattice constant mismatch and thermal budget constraints that otherwise present difficulties in heteroepitaxy-based methods. Theoretically, the process can be scaled up to reconcile wafer size mismatches between Si and non-Si materials through the use of over-patterning and superlattice designs on the source wafers to ultimately realize a commercial "step and transfer" style implementation. Experimentally, we have demonstrated the ability to transfer large arrays of high aspect ratio GaAs and InAs NRs to Si/SiO_2 substrates for use as heterogeneous channels in JL-FETs. The alignment accuracy and transfer yield are identified as the primary challenges to NTP-HGI; currently we have achieved an overlay error of ~16 μm and transfer yield < 10%, both of which could be improved with better tools and refinements to the NTP process. Mechanisms for yield loss were identified and suggestions for improvement thereof were proposed. Despite the successful transfer

of heterogeneous NRs, the fabricated JL-FETs did not show signs of electrical conduction, possibly because of poor contact formation or other problems originating from the transfer process. A deeper investigation is needed to identify the exact cause(s) for device failure.

To assess the potential of feature-level HGI in VLSI circuits, we also developed an evaluation framework to project the benefits in delay, power, and area afforded by 15nm InGaAs/Ge FinFET HGI over Si/Si non-HGI technology. The higher drive current and lower capacitance of InGaAs results in substantially reduced delay and power consumption compared to Si. More importantly, we directly mapped the effects of transfer misalignment into the design rule requirements for HGI circuit layouts. We showed that the use of post-transfer fin trimming will be mandatory to alleviate the problem of lateral misalignment-related area penalties in NTP-HGI. Despite additional area penalties caused by transfer misalignment, some block-level HGI designs actually consumed less total chip area because fewer buffers were needed to meet a given performance target. These findings give strong motivation to pursue HGI as a technology option to improve digital circuit performance for the nanoscale era beyond traditional scaling.

Finally, we analyzed the manufacturing costs of implementing HGI by either NTP or NHE in 22nm JL-FET technology using Si, Ge, and InGaAs as the material choices. We found that NTP-HGI incurs higher per wafer cost (up to 53% increase) compared to a non-HGI process, primarily because of the expensive substrates needed and the high cost of epitaxial growth. In comparison, NHE-HGI is substantially cheaper to implement because only Si substrates are needed. This suggests that NHE may be a commercially attractive option for HGI, although process-related challenges remain a concern.

7.7 Appendix I: Experimental Procedure for GaAs Transfer to SiO_2

Table 26. Process Flow for GaAs NR Transfer to SiO_2

Starting Materials	• 3-inch GaAs/$Al_{0.8}Ga_{0.2}As$/GaAs donor wafer grown by MBE as shown in Fig. 63. • SiO_2 receiving substrate, at least 1 cm × 1 cm • PDMS stamp, large enough to cover pattern size
GaAs Lithography (on full 3" wafer)	• Dehydration bake @ 200°C for 5 min • HDMS prime • Spin coat with SPR-700 1.2 resist @ 5000 rpm • Soft bake @ 95°C for 60 sec • Expose @ 140 mJ/cm^2 dose in ASML PAS 5500 stepper • PEB @ 115°C for 60 sec • Develop in AZ 300 MIF for 60 sec • Dice wafer into pieces for subsequent processing
GaAs Patterning	• Hard bake @ 100°C for 2 min • Dip in 37% HCl/H_2O (1:5) for 30 sec to remove native oxide • Etch GaAs in citric acid/H_2O_2 solution (20:1) for 60 sec • Remove PR in acetone
GaAs Undercutting	• Etch AlGaAs in dilute BOE solution (1 mL 6:1 BOE per 100 mL H_2O) to undercut NRs; required time depends on NR width and pattern density • Place PDMS on GaAs substrate and peel off quickly in direction parallel to NRs • Dip PDMS containing GaAs NRs in BOE to remove any backside AlGaAs residue • Dip SiO_2 substrate and PDMS containing GaAs NRs in H_2O_2 • Place PDMS on SiO_2 substrate and peel off slowly to transfer GaAs NRs • O_2 plasma clean SiO_2 substrate with GaAs @ 80W RF power, 50°C for 60 sec to remove hydrophobic PDMS residue
ALD Gate Dielectric	• Deposit 5 nm Al_2O_3 (50 cycles) by ALD @ 200°C
Gate Lithography	• Dehydration bake + HMDS prime • Spin coat with AZ NLOF 5510 resist @ 2500 rpm for 45 sec • Soft bake @ 90°C for 60 sec • Expose @ 120 mJ/cm^2 dose in Karl Suss MA6 contact aligner • PEB @ 110°C for 60 sec • Develop in AZ 300 MIF for 60 sec
Metal Gate Deposition	• Evaporate 100 nm Al at 25°C in CHA Mark 40 • Perform metal lift-off in acetone
S/D Lithography	• (same recipe as gate lithography) • Open contact windows in dilute BOE for 30 sec and dilute HCl for 30 sec
S/D Deposition	• Evaporate 50/25/50 nm AuGe/Ni/Au at 25°C in CHA Mark 40 • Perform metal lift-off in acetone
S/D RTA	• Rapid thermal anneal @ 400°C for 60 sec

7.8 Appendix II: Experimental Procedure for InAs Transfer to SiO_2

Table 27. Process Flow for InAs NR Transfer to SiO_2

Starting Materials	• 2-inch InAs/$Al_{0.4}Ga_{0.6}Sb$/InAs/GaSb donor wafer grown by MBE as shown in. • SiO_2 receiving substrate, at least 1 cm × 1 cm • PDMS stamp, large enough to cover pattern size
InAs Lithography (on full 2" wafer)	• Dehydration bake @ 180°C for 5 min • HDMS prime • Spin coat with PMMA 495A4 resist @ 3000 rpm • Soft bake @ 180°C for 60 sec • Expose @ 550 μC/cm^2 dose in Vistec EBPG 5000+ES electron beam writer • Develop in MIBK:IPA for 30 sec • Dice wafer into pieces for subsequent processing
InAs Patterning	• Hard bake @ 100°C for 2 min • Dip in 37% HCl/H_2O (1:5) for 30 sec to remove native oxide • Etch InAs in citric acid/H_2O_2 solution (20:1) for 60 sec • Remove PR in acetone
InAs Undercutting	• Etch AlGaSb in dilute NH_4OH solution (1 mL 29% NH_4OH per 10 mL H_2O) to undercut NRs; required time depends on NR width and pattern density • Place PDMS on InAs substrate and peel off quickly in direction parallel to NRs • Dip PDMS containing InAs NRs in BOE to remove any backside AlGaSb residue • Dip SiO_2 substrate and PDMS containing InAs NRs in H_2O_2 • Place PDMS on SiO_2 substrate and peel off slowly to transfer InAs NRs • O_2 plasma clean SiO_2 substrate with InAs @ 80W RF power, 50°C for 60 sec to remove hydrophobic PDMS residue
ALD Gate Dielectric	• Deposit 5 nm Al_2O_3 (50 cycles) by ALD @ 200°C
Gate Lithography	• Dehydration bake + HMDS prime • Spin coat with AZ NLOF 5510 resist @ 2500 rpm for 45 sec • Soft bake @ 90°C for 60 sec • Expose @ 120 mJ/cm^2 dose in Karl Suss MA6 contact aligner • PEB @ 110°C for 60 sec • Develop in AZ 300 MIF for 60 sec
Metal Gate Deposition	• Evaporate 200 nm Al at 25°C in CHA Mark 40 • Perform metal lift-off in acetone
S/D Lithography	• (same recipe as gate lithography) • Hard bake @ 150°C for 5 min • Open contact windows in dilute BOE for 30 sec and dilute HCl for 30 sec
S/D Deposition	• Evaporate 20/200 nm Ti/Al at 25°C in CHA Mark 40 • Perform metal lift-off in acetone
S/D RTA	• Anneal @ 350°C for 1 hr

Chapter 8

Supercapacitors for Microelectronics[25]

8.1 Background

Despite the ubiquity of dielectric capacitors in microelectronic circuits, their achievable capacitance per unit area (given by $C/A = \kappa\epsilon_o/d$, where κ is the dielectric constant, ϵ_o is the vacuum permittivity, and d is the dielectric thickness), is fundamentally limited by how thin the dielectric film can be made without incurring substantial leakage. For a planar metal-insulator-metal (MIM) capacitor with flat electrodes separated by a 1 nm SiO_2 dielectric (κ = 3.9), the theoretical maximum capacitance is ~35 fF/μm^2. High-k dielectrics such as HfO_2 ($\kappa \cong 20$) can extend this limit, but not by more than an order of magnitude. Conversely, scaling the dielectric thickness below 1 nm is generally undesirable because of the exponential rise in direct tunneling current. To circumvent these limits, researchers have made nanoporous dielectric capacitors which pack a greater electrode surface area into a smaller footprint area [163]. Among the highest reported areal capacitances were obtained for rolled-up nanomembranes (2.0 pF/μm^2) and single nanowires (1.4 pF/μm^2) [164], [165]. However, it is unclear whether such approaches will be scalable for practical on-chip use.

An alternative approach to improve areal capacitance is to use a different capacitance mechanism altogether. Electric double-layer capacitors (EDLCs), or "supercapacitors", store charge at the interface of a solid electrode and liquid electrolyte. For these devices, d is reduced to

[25] This chapter summarizes collaborative work with our group and Dr. Leland Smith, Jonathan Lau, and Prof. Bruce Dunn from the UCLA Department of Materials Science and Engineering. We are especially grateful for Dr. Smith and Mr. Lau's tireless efforts with the fabrication and characterization of the experimental devices and their invaluable contributions to this chapter.

approximately the size of the electrolyte molecule (several angstroms for ionic liquids) [166]. This results in higher areal capacitance for EDLCs compared to dielectric capacitors without the associated problem of higher tunneling current[26]. Furthermore the electrode is typically coated with a high surface area carbon layer: for an electrode having a 0.1 mg/cm^2 loading of carbon material with 1000 m^2/g surface area, the electrical surface area exceeds the apparent by a factor of 1000. Thus, large *A* is achieved in a small footprint area using carbon-coated electrodes. For these reasons, up to three orders of magnitude improvement in areal capacitance can be expected with EDLCs compared to traditional thin-film dielectric capacitors. The rate-limiting process in an EDLC is the rearrangement of ions at the electrode surface. In order for EDLCs to operate at high frequencies, it is necessary to engineer both the materials and architecture to facilitate fast ion transport.

EDLCs that are intended for bulk energy storage are typically made in a cylindrical format. The two electrodes are equidistant from one another and separated by liquid electrolyte. The cylindrical format, however, is difficult to fabricate on-chip at sub-mm^2 dimensions; instead, many groups are researching methods for fabricating on-chip EDLCs using a coplanar geometry. A diverse array of methods to form carbon-based electrodes have been reported including the pyrolysis of photoresist, stamping of carbon nanotubes and the laser scribing of graphene oxide [167]–[169]. To avoid leakage and evaporation issues associated with liquid electrolytes, many use some sort of ionic liquid immobilized in a silica matrix (ionogel) to confine the electrolyte and form solid state devices.

[26] Leakage current in an EDLC is governed by charge-transfer reactions at the electrode-electrolyte interface rather than quantum mechanical tunneling of free electrons through a potential barrier.

In this work, we report our efforts to use self-assembly as a scalable, room-temperature process for the selective deposition of carbon nanoparticles on gold electrodes for microscale carbon-ionogel EDLC integration on silicon substrates. We also show how choices in device geometry affect the bandwidth of such EDLCs. In order to guide the design of smaller, higher bandwidth EDLCs suitable for microelectronic circuit applications, we also develop a physical computer-aided design (CAD) model which can be used to simulate the performance of coplanar EDLCs at microscale dimensions. The novel fabrication process could enable future incorporation of silicon-compatible EDLC technology with microelectronic and nanoelectronic integrated circuits, while the simulation framework will be a useful tool for the device community to optimize EDLC designs for on-chip use.

8.2 Process Flow Template

A conceptual process flow for the integration of planar carbon-ionogel EDLCs on a silicon substrate is proposed in Fig. 82. Beginning with an oxide-covered silicon substrate (Step 1), a series of metal electrodes are deposited and patterned (Step 2). Next, an insulating hardmask ("well") layer is deposited and patterned to expose only the active electrode regions while covering the metal leads/pads. (Step 3). High surface area carbon material is then selectively deposited over the exposed electrode area without coating the surrounding oxide (Step 4), thus preventing electrical shorts. Once carbon has been deposited, ionic liquid electrolyte is cast over the exposed electrode area and solidified into ionogel (Step 5). Then, an insulating material is deposited over the ionogel to encapsulate the entire device (Step 6). Finally, interconnect vias are formed over the covered pads to connect the electrodes to neighboring readout circuits (Step 7).

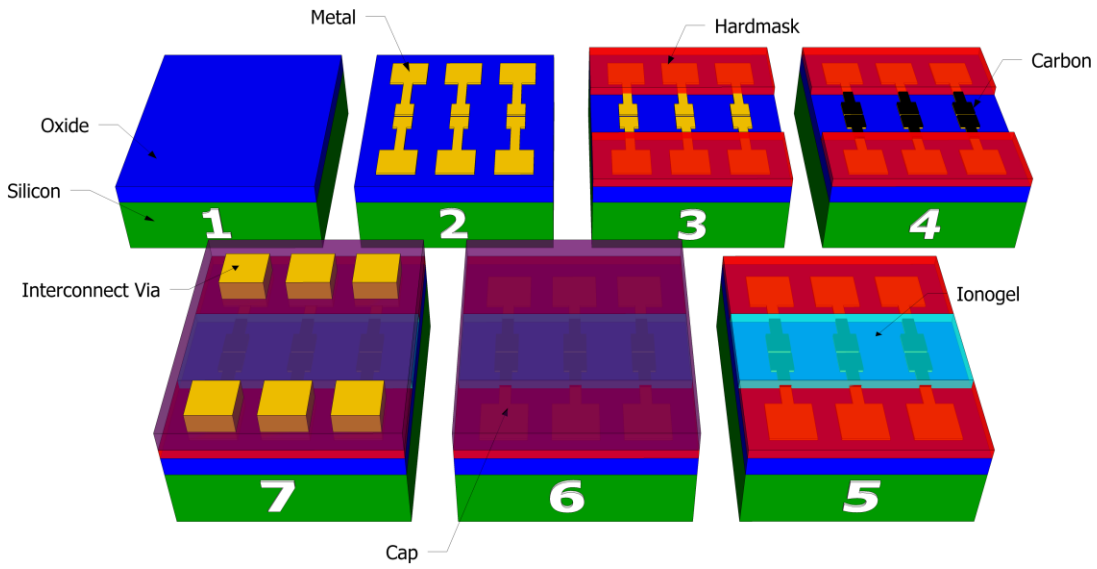


Fig. 82. Illustration of a generic process flow for integrating planar carbon-ionogel EDLCs on a silicon substrate. The numbers 1-7 indicate the sequence of processing steps.

In our experiments, we used 4" silicon (100) wafers with 1 μm thermally grown SiO_2 as the starting substrates. Thin films of Cr/Au (20 nm/100 nm) were evaporated at room temperature onto the substrates and patterned into electrodes with varying lengths, widths, and gap separations following metal liftoff. Multiple EDLC designs were explored in this work ranging from 1.54×1.1 mm^2 to 15×15 μm^2 in active electrode area and interelectrode gaps ranging from 400 μm down to 5 μm. Carbon nanoparticles were self-assembled onto the gold electrodes via a technique described in the experimental section. The electrolyte consisted of either: i) neat ionic liquid deposited by drop-casting, or ii) ionic liquid encased in ionogel by drop-casting or spin-coating, also detailed in the experimental section. For the EDLCs with hard mask wells, 1 μm Si_3N_4 was deposited by plasma enhanced chemical vapor deposition followed by photolithography and dry etching of the nitride to define the wells.

Due to time and resource limitations, not all of our experimental devices were subjected to the entire process sequence in Fig. 82. We emphasize that the intent behind Fig. 82 is merely to

propose a generic route to the eventual realization of on-chip integrated EDLCs. Thus, not all the steps shown in the figure (e.g., carbon deposition, ionogel synthesis, or well formation) were performed for the devices fabricated in this study. Rather the experimental results shown in this work represent the critical elements of the complete process. In this way, our work establishes the basis for achieving on-chip integrated EDLCs and highlights the need for continued work in the field.

8.3 Experimental Procedure

8.3.1 Fluoroalkylsilane (FAS) Treatment of Exposed SiO_2*:*

Carbon nanoparticles were selectively deposited onto patterned gold electrodes using self-assembly based on surface chemistry interactions. The SiO_2 surface on the silicon wafer was treated with a fluorinated alkylsilane, (tridecafluoro-1,1,2,2-tetrahydrooctyl)trichlorosilane (FAS) that suppresses the adherence of polar solvent molecules. The gold electrodes are inert to this chemistry. The FAS deposition method was adapted from a method reported by Jung et al [172].

Vapor-phase FAS functionalization was performed in a reaction flask (Ace Glass) equipped with an o-ring sealed joint. The head of the flask had two threaded ports connected to tubing with “Ace-Safe” thread-to-tube connections. One port was branched through two polytetrafluoroethylene (PTFE) needle valves (Omega) to a 25 mL vial and to room air for purging. The 25 mL vial was topped with a rubber septum so that the vial could be evacuated before introducing FAS via a syringe. The other port on the flask was connected by tubing to a liquid nitrogen trap, a pressure gauge (Omega), a ball valve (Swagelok) and finally a Welch 8905 vacuum pump in series. The ball valve allows the entire system to be isolated from the vacuum pump. This system achieved a base pressure of 40 mTorr.

Gold on silicon patterned electrodes were cleaned by sonicating for 10 minutes in a 1:1 mixture of isopropanol and acetone. The patterned wafers were then cleaned in Harrick PDC oxygen plasma cleaner for 8 minutes on high power. The plasma chamber was purged three times with oxygen and then pressure was set to 900 mTorr of oxygen while the plasma was on. After plasma cleaning, the samples were immediately transferred to the reaction flask. After loading the cleaned samples, the deposition chamber was evacuated for 10 minutes. The patterned wafers were exposed to three FAS-deposition cycles. Each cycle consisted of: i) 30 seconds of exposure to FAS vapor with the ball valve open (vacuum on), ii) three minutes of exposure to FAS with the ball valve closed, iii) six minutes of evacuation with the FAS valve closed and the ball valve open. After three FAS deposition cycles the reaction flask was purged three times with room air. The patterned wafers were then removed from the reaction flask and placed in a 60°C oven overnight.

8.3.2 Selective Carbon Deposition by Self-Assembly:

Carbon nanoparticle dispersions were prepared by dispersing Ketjen black (Printex XE-2B or KB) with a primary particle size of 35 nm in dimethylsulfoxide (DMSO). 1 mg of KB was added to 20 mL of DMSO and sonicated for 2 hours before use. The area around the patterned gold electrodes was masked using 40 μm thick PTFE tape, leaving an exposed area about 2 mm^2. To self-assemble the carbon electrodes, 40 μL of the KB dispersion was placed on the exposed portion of the FAS-functionalized silicon wafer. Next, a piece of celgard porous membrane was used to wipe away excess dispersion so that a smooth film of KB dispersion was seen to stretch uniformly across the exposed portion of the wafer. The electrode was placed on a 120°C hot plate so that the DMSO evaporated within 2 minutes. This process of drop-casting and evaporating the KB dispersion was repeated up to 40 times while the carbon coverage was evaluated using optical microscopy.

8.3.3 Ionogel Synthesis:

Ionogel was prepared according to the method reported by Membreno et al [173], [174]. The ionic liquid 1-butyl-3-methylimidazolium tetrafluoroborate ([BMIM][BF4]) was stored in an argon-filled glovebox to prevent the uptake of moisture. About 1 mL of [BMIM][BF4] was removed from the glovebox 30 min before the synthesis began. Sol was prepared containing a 2:2:5 volume ratio ratio of tetramethoxysilane : vinyltriethoxysilane : formic acid. This sol was stirred in a 39°C oil bath for 19 minutes after which [BMIM][BF4] ionic liquid was added resulting in 42% total ionic liquid volume in the sol. The sol/ionic liquid mixture was stirred for 30 seconds and then applied to the devices by both spin-coating and drop-casting. Spin-coated ionogel was applied to the bare B30 gold electrodes using an initial 1500 rpm (500 rpm/s ramp) spin for 1 minute, followed by either 3000 or 6000 rpm (1000 rpm/s ramp) for one minute. A Dektak 6 surface profilometer was used to measure the thickness of the spin-coated ionogel. The average thicknesses of the 3000 and 6000 rpm spin-coats were 3 μm and 1 μm, respectively.

Ionogel was applied to the KB electrodes by drop-casting. Prior to drop-casting ionogel, Mylar tape was used to create a 1 mm wide, 100 μm tall channel around the KB electrodes. Ionogel was applied by drop-casting and then spread with a razor blade, defining the thickness at 100 μm.

8.3.4 Electrochemical Characterization:

Cyclic voltammetry was performed using a Biologic VMP3 potentiostat. Impedance spectroscopy was performed using two setups. Impedance of the ionogel devices was tested using a Solartron 1287 potentiostat and 1252a function generator with a 10 mV amplitude, 0 V bias and range of 0.1 to 100 kHz. Impedance of the KB electrodes with liquid [BMIM][BF4] electrolyte (50 μL drop size) was measured under a probe station using an HP 4284A LCR meter with 10 mV ac

amplitude and 0 V bias from 20 Hz to 1 MHz. The measured impedance dispersions were represented by the series RC circuit model containing frequency-dependent capacitance and resistance [175].

8.4 Experimental Results and Discussion

8.4.1 Millimeter-Scale Gold-Ionogel EDLCs

We first evaluate the performance of EDLCs having ionogel electrolyte on planar gold electrodes without carbon coating. Because of their relative simplicity compared to carbon-based EDLCs, these devices are also more straightforward to model using the simulation framework that will be introduced in Section 8.5. Four different electrode geometries were investigated as shown in Table 28. Ionogel films were applied in either 1 or 3 μm layers to the gold electrodes as described in the experimental section.

Table 28. Specified Dimensions of "B30" Ionogel on Bare Gold Supercapacitors.

Device Name	Electrode Width (μm)	Electrode Length (μm)	Electrode Gap (μm)	Total Area (μm^2)
B31	2000	530	50	2.12×10^6
B32	2000	505	100	2.02×10^6
B33	2000	455	200	1.82×10^6
B34	2000	355	400	1.42×10^6

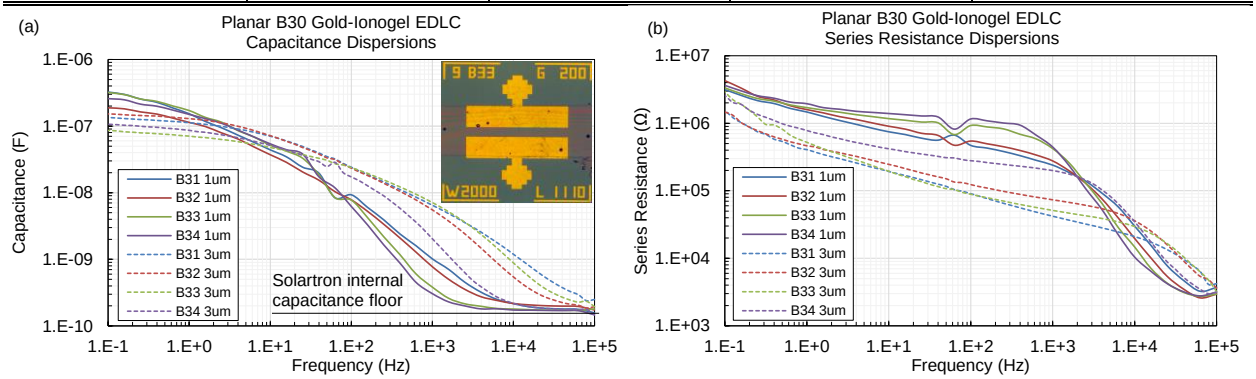


Fig. 83. Measured (a) capacitance and (b) series resistance dispersions for B30 gold-ionogel EDLCs. Solid (dashed) lines correspond to the 1 μm (3 μm) gel devices. The inset in (a) is an optical micrograph of one of the measured B33 devices. The color fringing in the ionogel is indicative of film thickness variations.

The capacitance and resistance dispersions as obtained from impedance spectroscopy for the 1 and 3 μm ionogel film B31–B34 devices are shown in Fig. 83[27]. Several trends are expected which we explain as follows. First, thicker gels (e.g., 3 μm vs. 1 μm) should yield less series resistance and greater capacitance at high frequencies. The reason for less resistance is the greater cross sectional area for diffusive flux of ions through the electrolyte in response to the ac signal; this is consistent with the concept of lower sheet resistance (equal to resistivity divided by film thickness) in thin solid films. Second, shorter electrode gaps should also yield less series resistance and higher capacitance for obvious reasons. At higher operating frequencies, the electrolyte resistance will bottleneck the response of the electric double-layer (EDL) to the ac signal, which will result in lower effective capacitance of the EDLC as the frequency rises. By reducing the total electrolyte resistance, either through the use of thicker ionogel films or shorter electrode gaps, we expect that the effective bandwidth of the EDLC will increase (i.e., we retain more of the peak double-layer capacitance at higher frequencies).

In the capacitance data of Fig. 83(a), at frequencies greater than 10 Hz the trends are evident: devices with thicker gels and smaller gaps give higher capacitance. The capacitance of these devices also drops below the 150 pF internal limit of the Solartron measurement system at higher frequencies more slowly compared to the samples with thinner gel and wider gaps. The one outlier from these trends is the B33 3 μm sample, which may have a geometrical or compositional defect in the ionogel coating. From Fig. 83(b) up to about 2 kHz the resistance trends are also evident: wider gap spacing and thinner ionogel layers give higher series resistance. At higher frequencies (above 1 kHz for the 1 μm ionogel devices), the series resistance appears to rapidly drop, however this also a consequence of the Solartron's 150 pF input capacitance and not indicative of true

[27] In this work we represent all measured impedance data by the first-order series RC representation, which assumes the impedance network consists of a single frequency-dependent resistor and capacitor connected in series.

behavior from the EDLC itself; a detailed explanation for this can be found in Appendix II. Finally, the noise around 60 Hz is due to interference from nearby electronic equipment.

At lower frequencies where the electrolyte resistance is no longer expected to limit the EDL response, there is no clear trend in the data. Surprisingly, the thinner gels appear to give higher capacitance below 1 Hz. A possible explanation is that the higher spin rate (6000 rpm vs. 3000 rpm) may have resulted in greater redistribution of ionic liquid away from the vicinity close to the electrode gap, thereby reducing the ionic concentration and the EDL coverage where it matters most. We should point out that the gel thicknesses reported here are average measured values obtained from surface profilometry. Variations in film thickness, as shown in the inset of Fig. 83(a) for one of the B33 devices, may be responsible for the non-monotonic capacitance trends at low frequencies. Because of experimental variations from device to device, it is premature to draw any conclusive trends from the low-frequency data.

8.4.2 Sub-Millimeter-Scale Carbon-Ionogel EDLCs

The insights from the experiments in Section 8.4.1 suggest that good high-frequency capacitance can be obtained from electrodes with small gaps and thicker electrolyte coatings. Furthermore, EDLCs require a high surface-area carbon electrode for maximum performance. With these design principles in mind, we fabricated a prototype on-chip EDLC with significantly smaller electrodes (100 μm wide) and interelectrode gaps (10 μm) which better represent the desired scale for microelectronics integration.

KB was applied to the gold electrodes using the self-assembly (SA) technique described in the experimental section. The improvement in capacitance from the addition of KB onto the gold

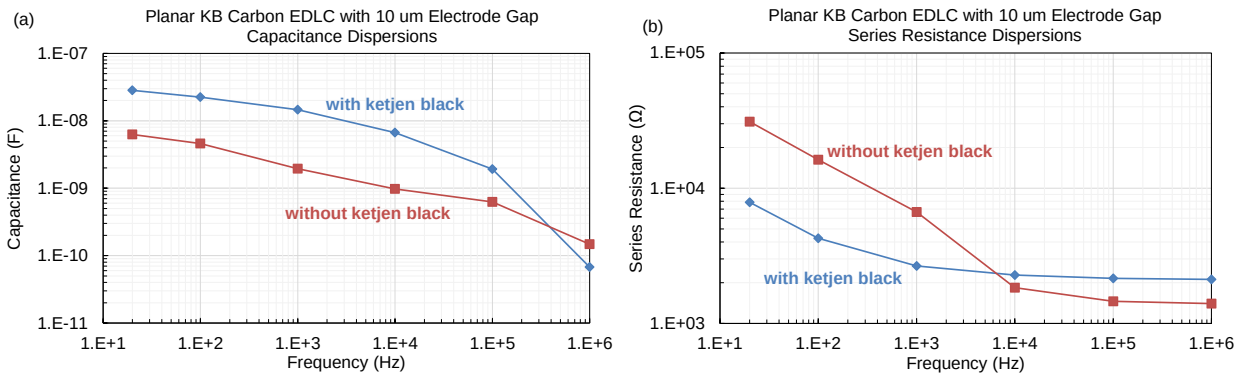


Fig. 84. Comparison of series capacitance and resistance dispersions for 10 µm gap supercapacitors with and without KB in neat ionic liquid.

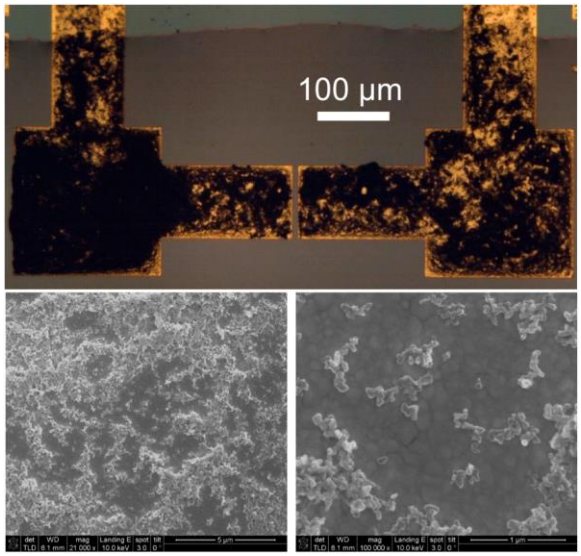


Fig. 85. (top) Optical microscope image of KB self-assembled electrodes coated with 100 µm thick ionogel. (bottom) SEM images of ketjen black particles self-assembled on the gold electrodes.

electrodes with 50 µL drops of neat [BMIM][BF4] electrolyte was measured by impedance spectroscopy using the HP LCR meter. As shown in Fig. 84, the capacitance is increased by almost an order of magnitude up to 100 kHz from the addition of KB.

Another set of KB electrodes was also prepared by SA but this time coated with a 100 µm thick ionogel layer by doctor-blading. One such device is shown in Fig. 85 and we see that the SA technique is able to carbon-coat the gold electrodes relatively uniformly without creating any electrical shorts across the 10 µm gap. Unfortunately, the ionogel ended up coating a large area (0.14 mm^2) of the gold electrodes, including portions which were not intended to be covered (i.e., the square pads and upper leads). Some of these regions are over 400 µm apart and are unlikely to

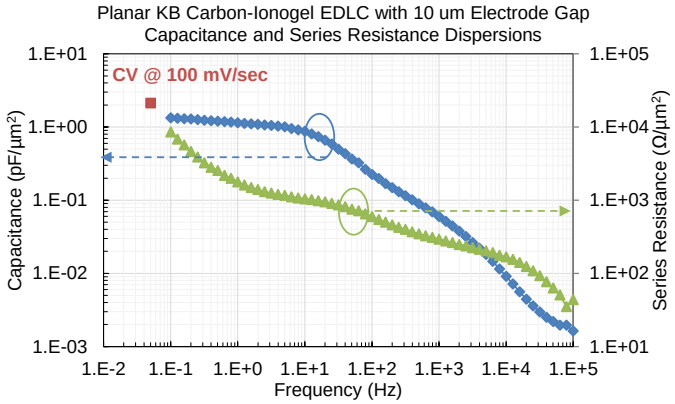


Fig. 86. Comparison of areal capacitance and series resistance measurements on the 10 μm gap carbon-ionogel supercapacitor.

contribute much capacitance at high frequencies. Further refinements, such as the incorporation of wells (see Section 8.4.3), should result in smaller device areas and significantly higher areal capacitance at high frequencies.

The capacitance of the device shown in Fig. 85 was measured by both impedance spectroscopy and cyclic voltammetry (CV) from 0 to 1 V at 100 mV/s. The low-frequency capacitance was found to be 297 and 185 nF by CV (at 0.05 Hz) and impedance spectroscopy (at 0.1 Hz), respectively. The charge and discharge capacity of the device measured by CV were 333 and 256 nC, respectively, suggesting that small irreversible faradaic processes are occurring at the electrode. For the most part, the capacitance values measured by impedance and CV are in relatively good agreement. In Fig. 86 the capacitance measured by these two methods is plotted normalized to the total electrode area covered by ionogel (0.14 mm^2). Capacitance values as high as 1 pF/um^2 are achieved for frequencies up to 10 Hz.

8.4.3 Micrometer-Scale Well EDLCs

While the KB/ionogel EDLC in Fig. 85 is small compared to many of the reports in the literature, on-chip EDLCs will have to be made even smaller in order to supplant dielectric capacitors. Fig. 87 shows some of our initial results for 15×15 μm well capacitors with areas defined by

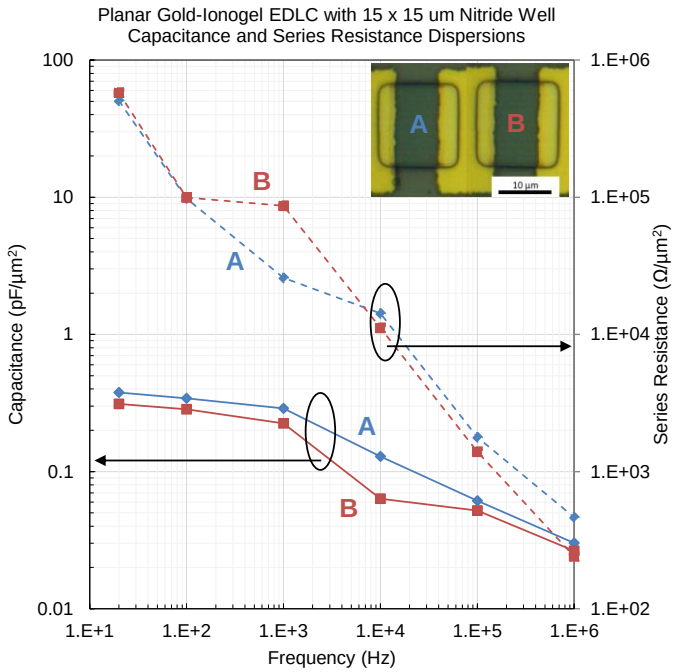


Fig. 87. Capacitance and series resistance dispersions for 15×15 μm well gold-ionogel supercapacitors. The inset shows optical micrographs of two such devices, denoted "A" and "B". The well openings are indicated by the black square outlines.

a 1 μm layer of Si_3N_4 and ionogel applied by spin-coating at 3000 rpm. Initially, we tried to deposit KB by SA on these devices as well, however we found that the deposition was not as selective and electrical shorts would form. For this reason, the results to be presented are for devices with bare gold electrodes only.

The impedances of two 15×15 μm well EDLCs, labeled "A" and "B" in Fig. 87 were measured under a probe station with the HP LCR meter. Because the ionogel only covers the gold areas exposed by the square well opening, these structures do not have the long diffusion paths seen in the device in Fig. 85. Despite not having carbon on the electrodes, they achieve almost an order of magnitude higher areal capacitance at 10 kHz compared to the device in Fig. 86. However, at low frequencies their capacitances saturate to less than half the value compared to the KB/ionogel device (0.3-0.4 vs. 1 pF/μm^2). The difference between the A and B curves is most likely related to how completely the ionogel filled the nitride wells in each case. We should point out that several

other samples prepared in the same way showed no measurable capacitance, suggesting that the ionogel has trouble completely filling the well. Because on-chip EDLCs will undoubtedly require well-defined active areas in order to ensure their performance is predictable and consistent from device to device, it is paramount that any and all problems related to incomplete coverage of the well by carbon and/or ionogel be addressed. Clearly, additional process refinements will be needed to bring the performance of these microscale EDLCs closer to their full potential, but the results so far are encouraging.

8.4.4 Benchmarking

Integrated on-chip EDLCs could offer unprecedented levels of areal capacitance that even the most sophisticated dielectric capacitors cannot, and for this reason many researchers are pursuing different methods to fabricate coplanar EDLCs. Table 29 summarizes some recent reports of on-chip EDLCs in the literature, which use various methods of selective carbon deposition. Our microscale devices reported here perform well in areal capacitance, especially considering the thinness of the electrodes. These devices are unique in their ability to deliver high areal capacitance in a very small size and with narrow electrode spacing.

Table 29. Benchmark Comparison against Other On-Chip Supercapacitors from Recent Literature.

Carbon Deposition Method	Device Area (mm^2)	Electrode Spacing (μm)	Electrode Thickness (μm)	Electrolyte	Capacitance by CV / sweep rate	Time Constant (s)	Reference
Pyrolysis of photoresist	22.5	300	1.2	Fumed silica and [EMIM][TFSI]	1.5 pF/μm^2 10 mV/s	0.1	Wang et al [167]
Chemical vapor deposition / stamping	6.4	35	50	Ionogel [EMIM][TFSI]	4.3 pF/μm^2 100 mV/s	0.021	Hsia et al [168]
Laser scribing graphene oxide	40	150	7.6	Poly(vinyl alcohol) and H_2SO_4	23 pF/μm^2 10 mV/s	0.033	El-Kady et al [169]
Electrophoretic deposition	25	100	7	1 M ET_4NBF_4 in PC	17 pF/μm^2 1 V/s	0.026	Pech et al [170]
Carbon sputtering	40	600	0.2	[BMIM][NTf2]	0.8 pF/μm^2 80 mV/s	3.0	Bettini et al [171]
KB self-assembly	**0.14**	**10**	**0.1**	**Ionogel [BMIM][BF4]**	**2.1 pF/μm^2 100 mV/s**	**0.12**	**This work**
No carbon used	**8×10^{-5}**	**5**	**0.1**	**Ionogel [BMIM][BF4]**	**0.3 pF/μm^2 *n/a***	**0.001**	**This work**

Besides areal capacitance, another important metric for the performance of EDLCs is the time constant ($\tau_o = 1/2\pi f_o$) where f_o is the frequency (i.e., bandwidth) at which the impedance phase angle reaches –45°. At frequencies above f_o the EDLC is unable to store/release charge efficiently due to diffusion resistance-limited behavior associated with sluggish ionic motion in response to the ac signal, while at frequencies below f_o the EDL can easily store/release charge in response to the signal. This figure of merit is especially important for on-chip capacitors designed for use in high-frequency applications traditionally reserved for dielectric capacitors. The time constant for our 0.14 mm^2 KB-ionogel device listed in Table 29 is somewhat long at 0.12 sec and will need improvement for practical use in electronic applications operating at the kHz range or faster, although we should still point out that the other EDLC demonstrations listed in Table 29 have not shown significantly better time constants either. Our 80 μm^2 gold-ionogel device, however, has a much better time constant at 1 ms which is encouraging despite the smaller areal capacitance at low frequencies.

There are other important considerations beyond what is listed in Table 29. The ideal manufacturing process for forming on-chip EDLC should be low-temperature, scalable to large areas, and compatible with typical foundry tools. These are requirements are especially important if EDLCs are to be heterogeneously integrated with complementary metal-oxide semiconductor (CMOS) circuits. Our SA technique of depositing commercially-available KB directly from solution has the advantage of being low temperature (T < 120°C) and scalable to wafer-level coverage if desired. We also showed that silica-based ionogel films can be used as the electrolyte which solves a major problem associated with packaging of ionic liquids or aqueous electrolytes in solid state EDLCs.

There are some remaining process issues to consider, however, such as the chemical stability of ionic liquids with certain metals. We observed that [BMIM][BF4] corrodes some common metals such as aluminum, titanium, and copper in the presence of water, while noble metals such as gold and platinum are more stable. Since gold is typically prohibited in a CMOS foundry environment, platinum could be substituted instead. Further work is also needed to determine if ionogel is compatible with dry etch and chemical vapor deposition processes which may be needed in the process template of Fig. 82. Regardless, the carbon-ionogel EDLC process developed in this work is a promising step forward.

Assuming the process-related challenges can be overcome and on-chip EDLC integration is successful, the next important question will be their design. To address this issue, we have developed a physical CAD model, described in the next section, which can be used for design exploration of the coplanar EDLCs described in the preceding sections.

8.5 EDLC Simulation and Modeling

To gain physical insight into how choices in EDLC design affect performance in terms of areal capacitance, resistance, and effective bandwidth when used as microelectronic circuit elements, it is imperative to develop a behavioral model for the EDLC which can be easily incorporated in device and circuit simulations. The underlying physics responsible for double-layer formation and Debye screening near a metal-electrolyte interface follow the Poisson-Boltzmann equation, which means we can employ technology computer-aided design (TCAD) simulations that are traditionally used for semiconductor devices to model the capacitive behavior of electric double-layers under certain assumptions with proper modifications [176], [177]. In this section,

we use Sentaurus TCAD [10] to generate structures resembling the carbon-ionogel EDLCs explored here and perform small signal ac simulations to investigate the effect of device scaling on frequency-dependent capacitance. The model is flexible enough to allow us to predict EDLC performance in response to changes in structural design (e.g., electrode sizes, gaps, layouts, etc.) as well as material parameters (e.g., electrolyte concentration, thickness, conductivity, etc.), and could be an invaluable asset to the device design community.

Before continuing, we should acknowledge a few limitations in the model. First, the ions are treated as infinitesimal point charges in the electrolyte and hence discrete crowding effects are neglected; this is the Gouy-Chapman model and is sufficient for a first-order model of EDL screening which is valid at low ion concentrations and potentials [166]. Later, we describe a method to account for ion size effects in describing the Stern layer. Second, EDLCs can show complex behaviors resulting from interfacial electrochemical reactions and diffusion kinetics through the electrolyte, especially at very low frequencies. These effects can be represented by charge-transfer and Warburg impedance elements in more complex equivalent circuits (e.g., Randles circuit) where they can be used as fitting parameters to match experimental data [177]. Unfortunately, it is difficult if not impossible to model such processes using physical RC elements from first principles, so we neglect these effects in our simulations. Third, we assume the electrolyte permittivity and conductivity are field-independent. In spite of these limitations, we will see that our purely electrostatic simulation model comes within an order of magnitude of matching experimental device behavior in terms of capacitive dispersion.

8.5.1 Simulation Details

Each of the simulated EDLC structures are composed of the following key regions/materials: i) two conductive electrodes, ii) a 'semiconductor' representing the Stern layer, and iii) another 'semiconductor' representing the electrolyte. We briefly explain each of these regions below.

In the device simulator, the electrolyte is modeled as an intrinsic semiconductor[28] where the electron/hole density of states, band gap, relative permittivity, and carrier mobilities are tuned to obtain: i) equilibrium electron/hole concentrations which are consistent with the anion/cation concentrations of the true electrolyte, and ii) a net conductivity that is consistent with that of the true electrolyte. In other words, the flow of electrons and holes in this (pseudo) semiconductor behaves as a proxy for the actual flow of anions and cations in the electrolyte. For all the results to follow, we simply assume the size and mobility of anions and cations are identical and thus the 'electron' and 'hole' properties are kept equal on all counts.

The Stern layer crudely accounts for the finite size of counter ions and is modeled as another semiconductor with a relative permittivity equal to that of the electrolyte and negligible electron/hole concentrations at equilibrium[29]. The thickness of the Stern layer is specified as the average radii of the anion and cation species. Because of the negligible free carrier densities, virtually no charge screening occurs in the Stern layer and hence it behaves as an insulator with large but finite resistance. We can then set the zero-bias resistivity of the Stern layer by adjusting the elec-

[28] The purpose of modeling the electrolyte as a (pseudo) semiconductor is to allow the electrostatic potential and carrier densities to be described via a self-consistent Poisson-Boltzmann solution in the device simulator. Obviously, the ionic electrolyte is not a 'real' semiconductor in any true sense, but for our purposes this is a satisfactory workaround.

[29] For convenience, the electron and hole density of states are adjusted to give equilibrium values of $n = p = 1$ cm^{-3} in the Stern layer which is also assigned a "band gap" of 9 eV.

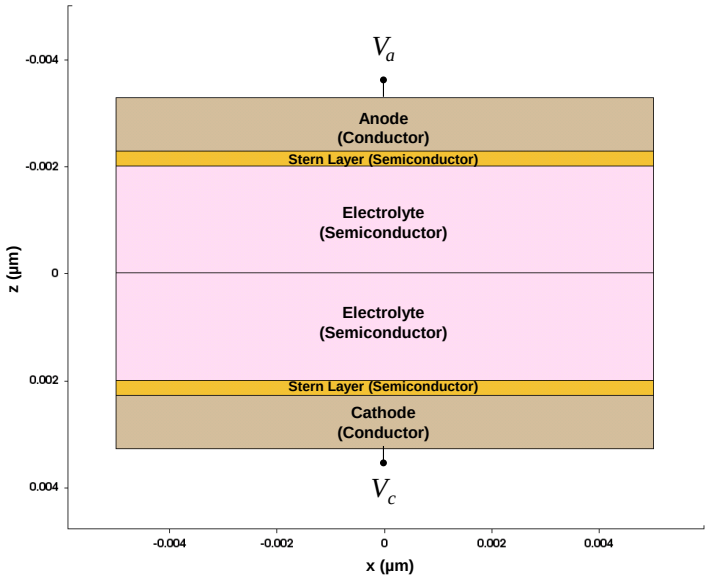


Fig. 88. Simulation model for the sandwich configuration EDLC.

tron/hole mobilities to yield the desired effective resistance across the Stern layer which is responsible for dc leakage[30]. Lastly, the metal electrodes are treated as ideal conductors with a work function chosen to obtain a zero flat band voltage.

Fig. 88 shows the structure used to simulate a 1-D vertical sandwich EDLC with 4 nm thick electrolyte, 1 nm thick electrodes, and a 2.8 Å thick Stern layer. The electrolyte represents the ionic liquid [BMIM][BF4] used in our experiments with anion/cation concentration of 3.22×10^{21} cm^{-3}, calculated from the known density (1.21 g/cm^{-3}) and molecular weight of the liquid (226 g/mol), and relative permittivity of 12.7. To save runtime, we can exploit symmetry in the setup of Fig. 88 by solving only one half of the device with the proper boundary condition at the $z = 0$ plane and simply divide the net capacitance by two for consistency with the full structure. Regardless of whether the full or half structure is simulated, we find it necessary to introduce an ohmic

[30] Since leakage only occurs in non-equilibrium, the actual Stern resistance will deviate somewhat from the desired value since $n, p \neq 1$ cm^{-3} under non-zero voltage conditions. However, we find that for small applied biases (0.1 V or less), n and p stay within 1 to 2 cm^{-3}, so the deviation in resistance will stay within an order of magnitude of the desired value.

contact at $z = 0$ with $V(0) = (V_a + V_c)/2$ to ensure the carrier densities return to equilibrium deep in the electrolyte bulk, where V_a and V_c are the anode and cathode potentials.

To verify that the simulation setup properly models double-layer screening near both metal-electrolyte interfaces, in Fig. 89 we show the potential and space charge distribution inside the sandwich EDLC when a steady-state voltage $V_{dc} = V_a - V_c$ up to 0.2 V is applied. As expected, the plots resemble those of back-to-back symmetric MOS capacitors except for the much thinner "depletion layer", which is now the diffuse layer in EDLCs. For ion concentration of 3.22×10^{21} cm^{-3}, the diffuse layer thickness (i.e., Debye length) is a mere 0.5 Å. Most of the potential is dropped across the Stern layers rather than the diffuse layers due to the high ion concentration, and, as expected, far from the diffuse layers there is no potential drop or net charge in the bulk electrolyte. We should note that such a thin (4 nm) electrolyte was simply chosen for visual clarity

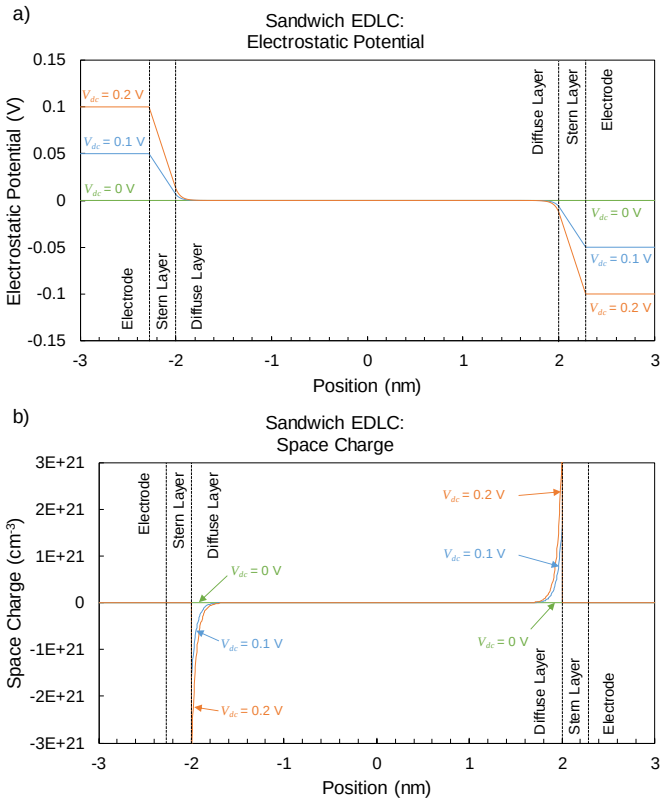


Fig. 89. Plots of (a) electrostatic potential and (b) space charge in a simulated sandwich EDLC for different DC voltage biases. The supercapacitor has a 4 nm thick electrolyte, 2.8 Å thick Stern layer, and 1 nm thick electrodes.

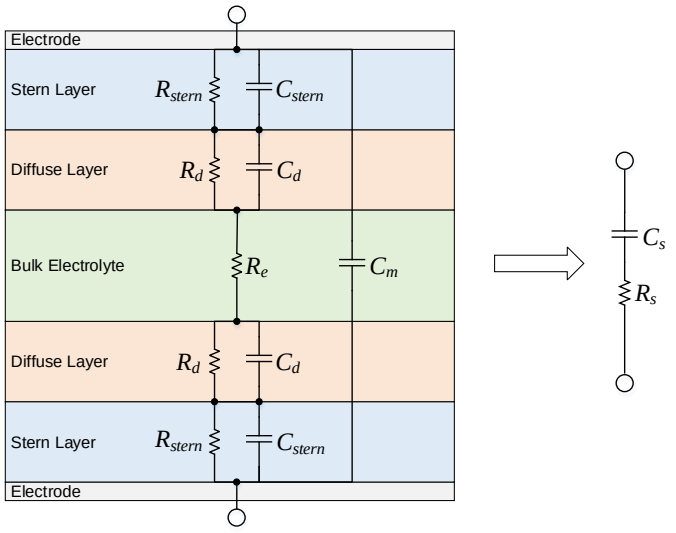


Fig. 90. Equivalent circuit diagram for the simulated EDLC model showing individual contributions from the Stern, diffuse, and bulk electrolyte regions.

of the sub-angstrom diffuse layers in Fig. 89, and that the same results would be obtained for a thicker electrolyte as well.

The important features of the EDLC can be represented using the equivalent circuit diagram shown in Fig. 90, which in turn can be modeled as a frequency dependent series RC circuit with a frequency dependent capacitance C_s and resistance R_s. Here, the total capacitance is composed of several elements: the individual capacitances across the Stern layer (C_{stern}), the diffuse layer (C_d), and the bulk electrolyte between the metal plates (C_m). Normally, both C_{stern} and $C_d \gg C_m$ so the net EDLC capacitance (at low frequencies) is half the double-layer capacitance given by $C_{dl} = (1/C_{stern} + 1/C_d)^{-1}$. When the electrolyte resistance R_e is large, however, the effective bandwidth of the double-layer capacitance C_{dl} drops, causing a reduction in the total capacitance C_s of the system at higher frequencies. This reduction may be interpreted as a case of ineffective counter-ion response to higher frequency signals due to poor ionic mobility (~10^{-6} cm^2/Vs). Eventually the capacitance reduces to that of a standard MIM capacitor (C_m) when the double-layer fails to respond to the ac signal altogether.

Additional resistive elements appear across the Stern (R_{stern}) and diffuse layers (R_d) which account for the finite parallel resistance and leakage current in the dc limit. R_{stern} models charge-transfer resistance at the electrode-electrolyte interface and can be adjusted (by tuning the Stern layer resistivity) to match experimental leakage measurements. R_d represents the resistance across the spatially varying diffuse layer and, in principle, should be modeled as a distributed RC network along with its counterpart C_d. Instead, we treat R_d and C_d as lumped elements for simplicity where the resistivities of the diffuse layer and bulk electrolyte are assumed to be the same. Formulas for each of the circuit elements are provided in Appendix I.

8.5.2 Coplanar EDLC Modeling

For the coplanar EDLC, we use the basic setup shown in Fig. 91 which consists of two electrodes with Stern layers in a side-by-side layout with an interposing electrolyte and an underlying oxide-covered substrate. Again, the indicated sizes for the different regions in Fig. 91 are for visual clarity purposes, but these can be varied. Unlike the sandwich EDLC, the planar structure

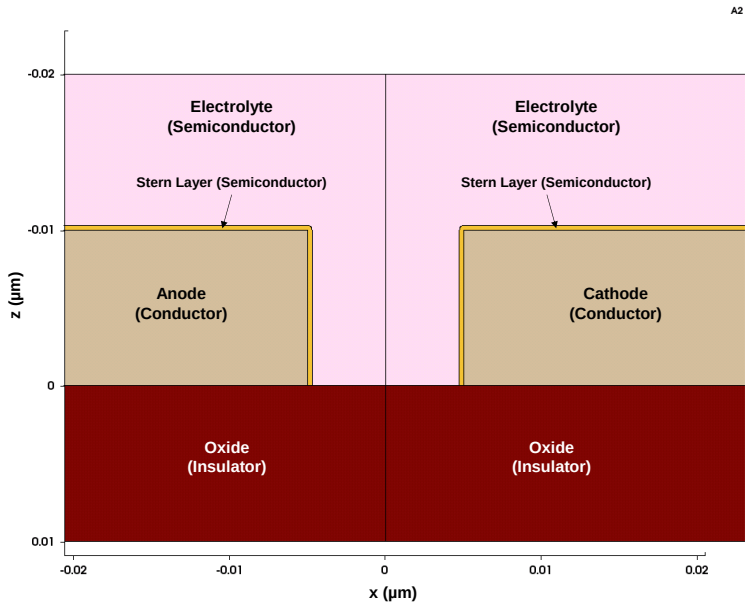


Fig. 91. Simulation model for the coplanar configuration EDLC.

is now 2-D which means analytical expressions cannot be obtained for the equivalent circuit elements in Fig. 90, so we must rely exclusively on TCAD simulations. Like the sandwich EDLC, we can exploit symmetry by solving only one half of the structure and imposing the proper boundary condition at the $x = 0$ plane, but only if the substrate potential is floating or fixed at $(V_a + V_c)/2$. If the substrate is held at any potential other than $(V_a + V_c)/2$, then symmetry is broken and the parasitic capacitance between the substrate and one electrode may be greater than the other.

To validate our TCAD model, we compared the simulated capacitance and series resistance dispersions in Fig. 92 with one of our experimental B31 EDLCs from Table 28. The device has 450 µm long and 2 mm wide electrodes, 200 µm gap spacing, and 1 µm ionogel. The simulations appear to underestimate the bandwidth by about 10× when compared to the measured data in Fig. 92(a), but the low-frequency capacitance and decay slope are on the correct order of magnitude. It is possible that the ionogel conductivity was somewhat higher that what we assumed (1 mS/cm) based on our previous devices, possibly due to hygroscopic absorption during storage and/or testing. Some reported values for [BMIM][BF4] conductivity at room temperature are in the range of 3 to 4 mS/cm, and it is known that ionic liquid conductivity tends to increase with frequency and temperature, so these effects may also partially explain the discrepancy [178], [180]. The experimental high-frequency capacitance saturates at 150 pF and is due to the parasitic input capacitance

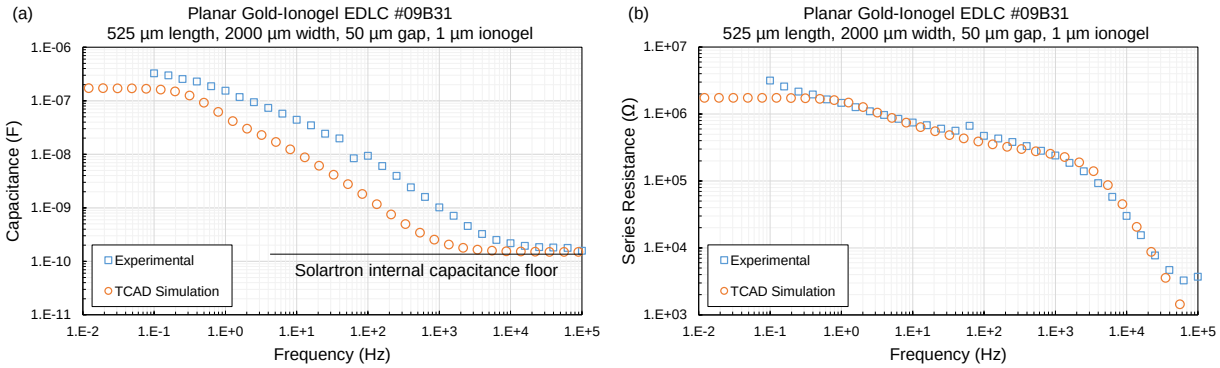


Fig. 92. Comparison of experimental planar EDLC capacitance and series resistance versus simulated values from the TCAD setup.

from the Solartron measurement system; this minimum capacitance has been added to the simulated results for comparison purposes only but it does *not* represent any part of the intrinsic device behavior. From Fig. 92(b), the series resistance matches quite well with experiment in both devices. For the simulated device, an infinite Stern resistivity has been assumed in the model based on the very high parallel resistance (about 1 GΩ) extracted from dc leakage measurements in open circuit conditions. Since the parallel resistance (i.e. the Stern resistance) is many orders of magnitude larger than that of the electrolyte, it has negligible influence on the dispersions of Fig. 92. Overall, the simple physics-based TCAD simulation is quite reasonable considering it does not rely on arbitrary parameter fitting; i.e., the values for all important parameters in the model are based on, or adjusted to be consistent with, the real electrolyte material.

In Fig. 93, we explore the effect of varying the electrode length, electrode gap, and electrolyte thickness on the resulting capacitance dispersion for planar EDLCs. The baseline design is

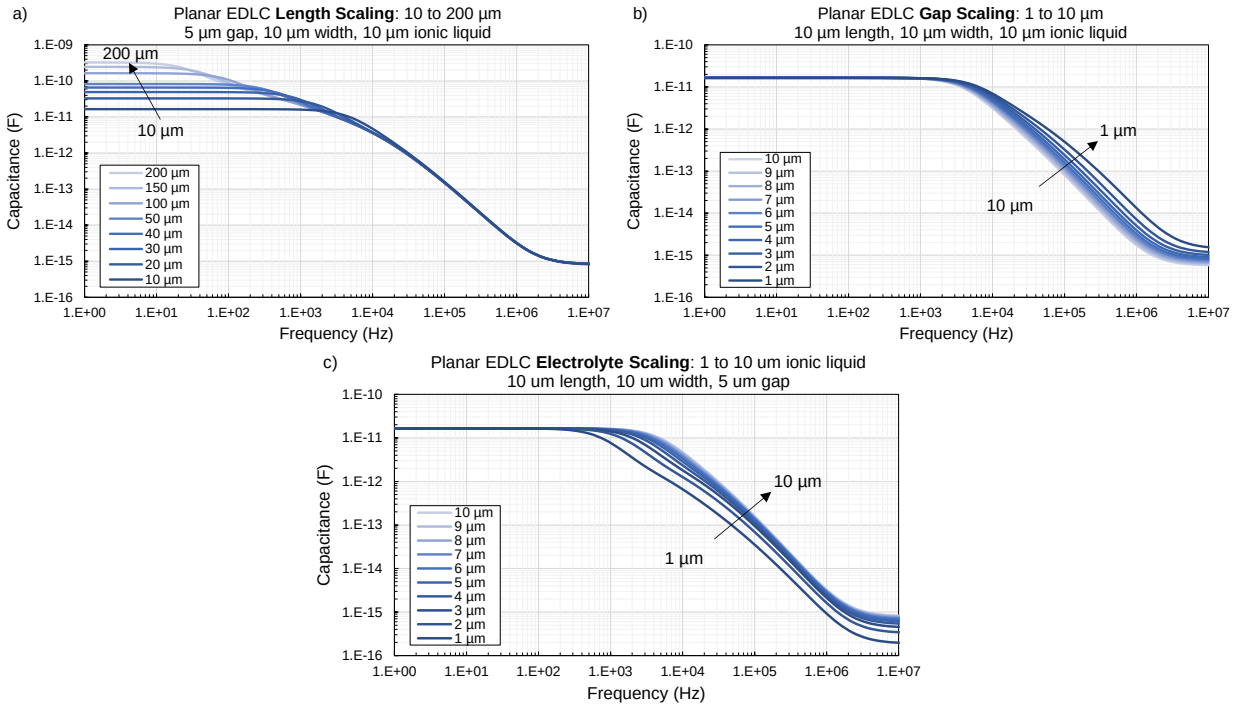


Fig. 93. Effect of scaling the (a) electrode length, (b) gap distance, and (c) electrolyte thickness on the capacitance dispersion of planar EDLCs. The electrolyte is assumed to be [BMIM][BF4] ionic liquid with 1 mS/cm conductivity and 2.8 Å Stern layer.

for a device with 10 μm long and wide electrodes, 1 μm gap distance, and 10 μm thick ionic liquid electrolyte. When the electrodes are elongated from 10 to 200 μm, a linear increase in the dc capacitance is observed in Fig. 93(a) because of more EDL coverage; however the increase in capacitance is only beneficial so long as there is enough bandwidth to harness it. In other words, every additional segment ΔL of electrode area must connect through a longer path of (resistive) electrolyte to the opposite electrode, ultimately reducing the bandwidth of the extra double-layer coverage. Interdigitated designs would likely mitigate this problem, although the lower EDL coverage per device footprint area may limit the gains of using long interdigitated fingers compared to wide blocks.

When the gap distance is increased from 1 to 10 μm, there is a noticeable loss in bandwidth as shown in Fig. 93(b). This is caused by larger electrolyte resistance between the electrodes when the gap increases. We see a similar trend in Fig. 93(c) when the electrolyte thickness is reduced from 10–1 μm, since there is more congestion of ionic transport (i.e., current crowding) through

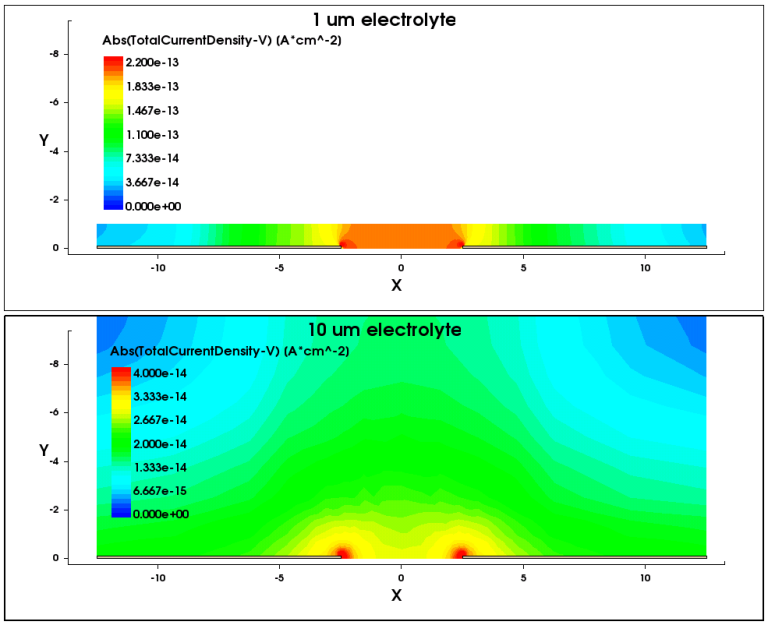


Fig. 94. Spatial current distributions in coplanar EDLCs with 1 um and 10 um thick electrolyte films with a 100 mV applied voltage. Crowding effects are more visible in thinner electrolytes.

the thinner film which results in higher resistance; this effect is illustrated in Fig. 94. Thus, to increase EDLC bandwidth, designs should aim to minimize the electrode gap separation and allow for sufficiently thick electrolyte to prevent a resistance bottleneck. Recall that these trends were also observed in the experimental B30 device results from Fig. 83 as well. Overall, the resistance of the electrolyte plays a significant role and designers should aim to reduce it as much as possible through materials engineering and/or design optimization.

8.6 Summary

We successfully demonstrated on-chip carbon-ionogel EDLCs at microscale dimensions on silicon substrates. The evaporation induced self-assembly technique allowed KB to be selectively deposited on the electrode surfaces which enhances areal capacitance by up to ten-fold over a bandwidth of 10 kHz. Thin film ionogel consisting of [BMIM][BF4] ionic liquid embedded in a porous silica matrix confined the electrolyte for use in solid-state device applications. High areal capacitance exceeding 1 pF/μm^2 up to 10 Hz was achieved for our coplanar EDLCs with 10 μm gap spacing and 0.14 mm^2 electrode area. Our smallest 15×15 μm well devices achieved 0.3 – 0.4 pF/μm^2 capacitance up to 1 kHz with 5 μm gap spacing and 80 μm^2 electrode area. These results are among the best reported values for on-chip EDLCs at sub-mm^2 dimensions. Our CAD model for simulating EDLCs allowed us to explore how choices in EDLC design affect overall performance in terms of capacitance, resistance, and bandwidth. The physical model showed reasonably good agreement with experimental data from the coplanar gold-ionogel EDLCs without relying on fitting parameters. Through experiments and simulations, several key elements for improving

EDLC performance were identified such as reduction of the electrode spacing and electrolyte resistance to extend the EDLC bandwidth. Our findings will hopefully open a new pathway to realizing on-chip EDLC technology for general microelectronic applications beyond energy storage.

8.7 Appendix I: Formulas for EDLC Circuit Elements

Each of the passive elements in the equivalent circuit model for the sandwich EDLC in Fig. 90 can be expressed using the following equations:

$$C_{stern} = \frac{\epsilon_e A}{t_{stern}} \tag{17}$$

$$C_m = \frac{\epsilon_e A}{t_e} \tag{18}$$

$$C_d = \frac{\epsilon_e A}{\lambda_d} = A\sqrt{\frac{2\epsilon_e q^2 c_{\pm}}{kT}} \tag{19}$$

$$R_e = \frac{t_e}{A\sigma_e} = \frac{t_e}{2Aqc_{\pm}\mu} \tag{20}$$

$$R_d = \frac{\lambda_d}{A\sigma_e} = \frac{\lambda_d}{2Aqc_{\pm}\mu} \tag{21}$$

$$R_{stern} = \rho_{stern}\frac{t_{stern}}{A} \tag{22}$$

where ϵ_e is the electrolyte permittivity, A is the device area, λ_d is the Debye screening length, q is the electron charge, kT is the thermal energy, $c_{\pm}$ is the ion concentration, σ_e is the electrolyte conductivity, μ is the average ionic mobility, t_{stern} is the Stern layer thickness, t_e is the electrolyte thickness, and ρ_{stern} is the stern layer resistivity.

8.8 Appendix II: Explanation for Series Resistance in Fig. 83(b)

For the 1 μm ionogel B31-B34 EDLCs, the sudden drop in series resistance between 3-4 kHz directly coincides with the frequency range where the true EDLC capacitance falls below 150 pF (the input capacitance of the Solartron measurement tool). At this point, the actual impedance

network resembles the equivalent circuit of Fig. 90, with the exception that the net parallel capacitance on the right-hand branch (what was C_m originally) is now given by the parallel combination of C_m and $C_{solartron}$ = 150 pF. At f > 3 kHz, the 150 pF term dominates that parallel combination. On the left-hand branch, C_s and C_d have already begun to short out even at f > 1 Hz (notice the steady drop in capacitance with rising frequency), which results in a series RC path on the left branch that becomes less and less reactive (i.e., the phase angle approaches 0°) at higher frequencies. At this point, the true impedance resembles a simple parallel RC network more than it does a series RC network. However, since we are still using the series RC representation for the entire measurement range, the extracted "series resistance" appears to drop at higher frequencies, even though at no point are any of the intrinsic resistance elements changing to first order. Rather, it is only the capacitance of the left-hand branch of Fig. 90 which is changing, but because of the way our measured impedance is represented, that effect happens to manifest in the extracted "series

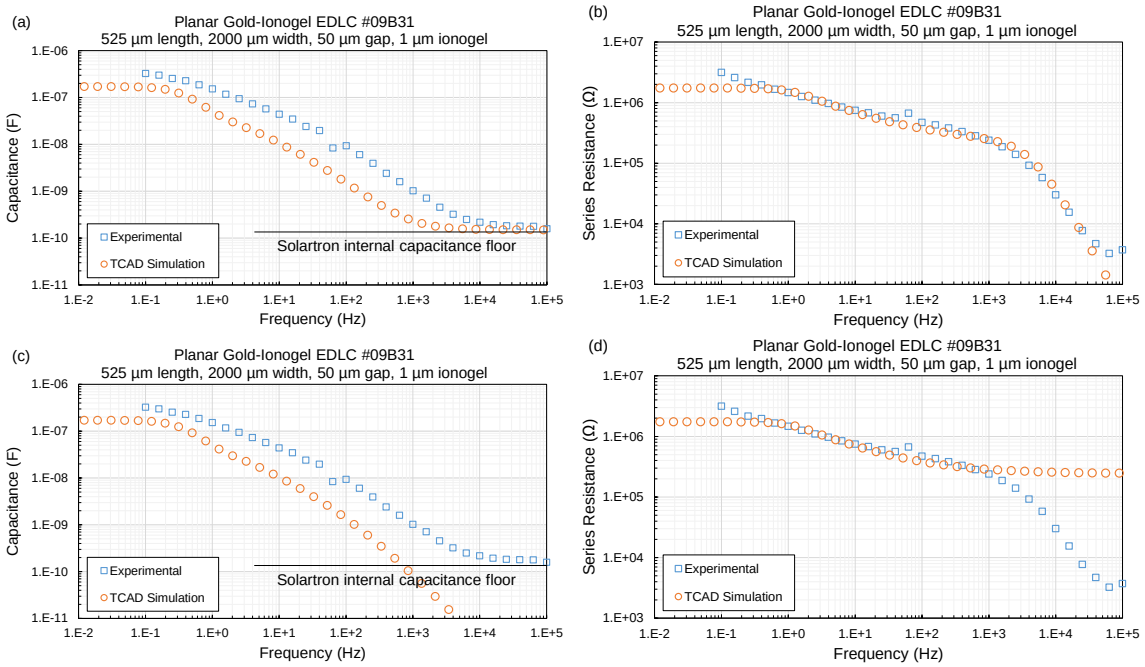


Fig. 95. Comparison of experimental planar EDLC capacitance and series resistance versus simulated values from the TCAD setup (a)–(b) with the addition of a fixed 150 pF capacitance to the simulated results for consistency with the experimental measurements. In (c)–(d), the extra 150 pF capacitor is removed, demonstrating the series resistance drop at f > 3 kHz is introduced by the Solartron.

resistance". To demonstrate this effect, in Fig. 95 we show that adding a fixed 150 pF parallel capacitance to the simulated B31 device characteristics reproduces the observed drop in series resistance, and removing the 150 pF capacitance instead reveals the true EDLC characteristics we normally expect to see (no sudden drop in series resistance at 3 kHz).

Now, the reader may wonder why we bother to use the series RC representation in the first place when reporting the impedance data (as opposed to using the parallel RC representation or even simply reporting the real and imaginary components Z' and Z" directly). The answer is simple: first, it is easier (at first glance) to interpret the capacitance of an EDLC represented by a series or parallel RC representation; second, the series resistance of the electrolyte is the most important factor that limits EDLC performance at high frequencies and hence it is the more interesting feature to observe and record, given the targeted application space for our proposed integration technology.

Chapter 9

Conclusion

At the beginning of this dissertation, several questions were posed concerning the potential of emerging technologies such as Si-based FinFETs, JL-FETs, TFETs, III-V FETs and heterogeneous integration of different devices to overcome the limits of pure Si CMOS. By now, we have hopefully addressed these questions—at least partially—in a manner which can help direct future research and development efforts by the semiconductor industry in choosing technology options with the best future outlook.

Device variability from LER and RDF was shown to be significant for nanoscale Si FETs especially at the 15nm node, and will undoubtedly become worse for future technology nodes. Some transistor architectures were revealed to be especially vulnerable to process variations, especially JL-FETs due to their fundamental nature as depletion-mode devices. Ironically, JL-FETs seem inherently less manufacturable compared to IM-FinFETs from a variability standpoint despite their allegedly easier fabrication thanks to their junctionless nature. IM-FinFETs on the other hand, suffer much less from LER and RDF and could remain manufacturable for several more generations. TFETs are more affected by LER and RDF than IM-FinFETs but less so than JL-FETs. We also found that for IM-FinFETs and TFETs, LER and RDF must be modeled simultaneously in order to get an accurate portrayal of the net variability due to LER-RDF interactions, especially those localized at the source-channel junction; for JL-FETs, however, this is not the case and the two variability mechanisms can be treated independently.

Interestingly, we saw that *n*-InGaAs JL-FETs actually show less vulnerability to RDF when compared to their Si counterparts thanks to suppressed carrier response in degenerately doped

materials. Along with much better nominal performance (i.e., higher drive current and lower capacitance), this makes a compelling argument for pursuing materials like *n*-InGaAs and *p*-Ge for heterogeneous integration and migrating away from Si altogether in future IC designs. We showed through experiments that NTP is a possible method to realize feature-level HGI of III-V materials on Si substrates, but that alignment accuracy and transfer yield would be significant challenges to the technology becoming truly feasible. Based on our conceptual vision of a foundry-scalable NTP-HGI process and considering the implications of NTP on design rules for circuit layouts, we showed that InGaAs/Ge HGI circuits can show sizeable improvements in speed, power, and even chip area compared to non-HGI circuits despite additional area overheads from transfer misalignment. On the other hand, the cost overhead of implementing NTP-HGI may be substantial compared to NHE-HGI.

In addition to studying new and different transistor technologies, we also demonstrated how microscale carbon-ionogel supercapacitors can be fabricated on silicon substrates for eventual heterogeneous integration with CMOS circuits. Our process is noteworthy because it is scalable, low temperature, compatible with traditional semiconductor processing, and solves the issue of packaging liquid electrolytes for use in solid state devices. The importance of minimizing the electrolyte resistance either through materials engineering or electrode design was underlined from our experimental findings and TCAD simulations of coplanar supercapacitors. Increasing the EDL bandwidth to at least the MHz range will be paramount for supercapacitor technology to be successful in high-frequency electronic applications where dispersion must be kept to a minimum—a role in which dielectric capacitors stand unmatched.

Still, there are many opportunities for further work in the fields of variability modeling for emerging devices and heterogeneous integration for future technologies which are worth mentioning. I should first point out that most of the suggestions to be presented are inspired from various unresolved problems that my colleagues and I faced while working on the different projects that eventually culminated into this dissertation.

First, concerning RDF and the Sano model in particular, a key unsettled question is how to properly (i.e., rigorously) choose the cutoff wavevector k_c. As we mentioned in Section 6.4, there is no clear answer at this point in time and yet the chosen value will directly scale the amount of RDF variability that will be observed in a given device. This limitation makes any and all estimates of RDF variability somewhat arbitrary if performed in a semiclassical (non-atomistic) environment, which is unfortunately a necessary compromise if true quantum mechanical simulations are impractical. Nevertheless, if the cutoff parameter can somehow be calibrated to deliver "correct" potential and carrier density fluctuations around individual or localized clusters of dopants, at least when compared to a true quantum mechanical solution such as one obtained from NEGF, this may serve as a better (but still imperfect) justification for choosing k_c.

Second, we have not investigated whether LER will be more or less significant for InGaAs FETs compared to Si. While in principle the same methodologies could be adapted from Chapter 2 through Chapter 6 to study this, some form of quantum corrections will likely be needed for InGaAs in the TCAD simulations, especially at the 15nm node and beyond. Local variations in the channel thickness from LER would almost certainly result in stronger quantum fluctuations for *n*-InGaAs because of its lower electron effective mass, but this could be difficult (if not impossible) to model accurately using the DGA with constant fixed parameters. Solving the 1-D Schrodinger equation in Sentaurus is an option, however it may be numerically difficult to use in terms of set

up and runtime. On the other hand, it is possible that spatially-dependent quantum corrections may be insignificant for the devices under study and that the DGA may still be sufficient, but we cannot prove this without first performing rigorous comparisons with quantum simulations from NEGF, for example. Regardless of how this problem is solved, the matter of LER's impact on III-V vs. Si FETs will need to be addressed in order to more fully understand how viable III-V transistors will be for future generations.

Third, there are many unexplored directions for the HGI evaluation framework that we did not have time to pursue here. The framework is general enough to allow for the treatment of other HGI processes besides coplanar transfer; examples of other possibilities include non-coplanar transfer (i.e., vertical integration) and NHE. Vertical integration could be achieved by sequentially transfer printing heterogeneous materials on different planes, one on top of another. This is akin to wafer bonding, except the bonded "substrate" is replaced with a soft stamp containing heterogeneous features to be transferred. The HGI interconnections would then be made using vias between the two layers to form 3D-HGICs which could potentially save more layout area compared to side-by-side HGI. On the other hand, NHE-HGI could also be studied using the framework assuming some known modifications to the design rules that may or may not be needed for in-grown channels. Ultimately, along with the work we have already presented on NTP-HGI, a comprehensive evaluation of the projected benefits of different approaches to realize HGICs would be highly useful for industry to identify the best option(s) for commercial adoption.

Lastly, regarding the future of on-chip supercapacitors, we proposed a fabrication strategy which could eventually lead to monolithic device integration with CMOS circuits; however we were unable to demonstrate a fully packaged and integrated device due to time and resource limitations. As we mentioned in Chapter 8, some remaining issues need to be addressed to close the

loop on our process scheme, but I do not believe these will be major hurdles to the success of this technology. Rather, I believe the biggest challenges will involve how to maximize the reliability and operating lifetime of these devices and how to minimize their performance dispersion compared to what traditional capacitors can deliver. And while considerable effort was made to study the small-signal ac response of our supercapacitors using impedance spectroscopy, their respective large-signal and transient responses were less studied. For signal processing applications that require fast charging/discharging times across a wide voltage range, any voltage-dependent capacitance in the EDLCs will also need to be well understood. In addition, some of our devices exhibited signs of dielectric retention and interesting open circuit leakage behaviors that were not covered in this dissertation, but definitely warrant further investigation. Clearly, there is much more work to do before on-chip supercapacitors can be relied upon to fill the role of dielectric capacitors in integrated circuits, but the research fronts are certainly pointed in the right direction.

References

[1] G. E. Moore, “Cramming more components onto integrated circuits,” *Electronics*, vol. 38, no. 8, pp. 114-117, Apr. 1965.

[2] C. Auth, A. Cappellani, J.-S. Chun, A. Dalis, A. Davis, T. Ghani, G. Glass, T. Glassman, M. Harper, M. Hattendorf, P. Hentges, S. Jaloviar, S. Joshi, J. Klaus, K. Kuhn, D. Lavric, M. Lu, H. Mariappan, K. Mistry, B. Norris, N. Rahhal-orabi, P. Ranade, J. Sandford, L. Shifren, V. Souw, K. Tone, F. Tambwe, A. Thompson, D. Towner, T. Troeger, P. Vandervoorn, C. Wallace, J. Wiedemer, and C. Wiegand, “45nm high-k + metal gate strain-enhanced transistors,” in *Proc. VLSI Symp.*, Jun. 2008, pp. 128-129.

[3] Y-K. Choi, K. Asano, N. Lindert, V. Subramanian, T-J. King, J. Bokor, and C. Hu, “Ultrathin-body SOI MOSFET for deep-sub-tenth micron era,” *IEEE Electron Device Lett.*, vol. 21, no. 5, pp. 254-255, May 2000.

[4] S. Thompson, M. Armstrong, C. Auth, M. Alavi, M. Buehler, R. Chau, S. Cea, T. Ghani, G. Glass, T. Hoffman, C-H. Jan, C. Kenyon, J. Klaus, K. Kuhn, Z. Ma, B. Mcintyre, K. Mistry, A. Murthy, B. Obradovic, R. Nagisetty, P. Nguyen, S. Sivakumar, R. Shaheed, L. Shifren, B. Tufts, S. Tyagi, M. Bohr, and Y. El-Mansy, “A 90-nm logic technology featuring strained-silicon,” *IEEE Trans. Electron Dev.*, vol. 51, no. 11, pp. 1790-1797, Nov. 2004.

[5] “Intel 22nm 3-D tri-gate transistor technology”, Intel Press Release, May. 2011.

[6] D. Hisamoto, L. Wen-Chin, J. Kedzierski, H. Takeuchi, K. Asano, C. Kuo, E. Anderson, K. Tsu-Jae, J. Bokor, and H. Chenming, "FinFET-a self-aligned double-gate MOSFET scalable to 20 nm," *IEEE Trans. Electron Devices*, vol. 47, pp. 2320-2325, 2000.

[7] L. L. Chang, Y. K. Choi, D. W. Ha, P. Ranade, S. Y. Xiong, J. Bokor, C. M. Hu, and T. J. King, "Extremely scaled silicon nano-CMOS devices," in *Proc. IEEE*, vol. 91, pp. 1860-1873, 2003.

[8] A. Burenkov, and J. Lorenz, “Corner effect in double and triple gate FinFETs,” *Proc. ESSDERC*, 2003, pp. 135-138.

[9] W. Xiong, “Multigate MOSFET technology,” *FinFETs and Other Multi-Gate Transistors* (edited by J. P. Colinge), Springer Publishing, 2008.

[10] *Sentaurus TCAD User Manual*, Synopsys, Inc., Mountain View, CA, Version F-2011.09.

[11] *International Technology Roadmap for Semiconductors*, 2013.

[12] O. M. Nayfeh and D. A. Antoniadis, "Calibrated hydrodynamic simulation of deeply-scaled well-tempered nanowire field effect transistors," in *Proc. SISPAD*, 2007, pp. 305-308.

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

[14] G. Masetti, M. Severi, and S. Solmi, “Modeling of carrier mobility against carrier concentration in arsenic-, phosphorus-, and boron-doped silicon,” *IEEE Trans. Electron Devices*, vol. 30, no. 7, pp. 764–769, Jul. 1983.

[15] C. Lombardi, S. Manzini, A. Saporito, and M. Vanzi, “A Physically based mobility model for numerical simulation of nonplanar devices,” *IEEE Trans. Computer-Aided Design*, vol. 7, no. 11, pp. 1164-1171, Nov. 1988.

[16] C. Canali, G. Majni, R. Minder, and G. Ottaviani, “Electron and hole drift velocity measurements in silicon and their empirical relation to electric field and temperature,” *IEEE Trans. Electron Devices*, vol. 22, no. 11, pp. 1045–1047, Nov. 1975.

[17] A. Asenov, S. Kaya, and A. R. Brown, "Intrinsic parameter fluctuations in decananometer MOSFETs introduced by gate line edge roughness," *IEEE Trans. Electron Devices*, vol. 50, pp. 1254-1260, 2003.

[18] W. Xiong, G. Gebara, J. Zaman, M. Gostkowski, B. Nguyen, G. Smith, D. Lewis, C. R. Cleavelin, R. Wise, Y. Shaofeng, M. Pas, T.-J. King, and J. P. Colinge, "Improvement of FinFET electrical characteristics by hydrogen annealing," *IEEE Electron Device Lett*., vol. 25, pp. 541-543, 2004.

[19] Y.-K. Choi, L. Chang, P. Ranade, J.-S. Lee, D. Ha, S. Balasubramanian, A. Agarwal, M. Ameen, T.-J. King, and J. Bokor, "FinFET process refinements for improved mobility and gate work function engineering," in *IEDM Tech. Dig*., 2002, pp. 259-262.

[20] A. Yahata, S. Urano, T. Inoue, and T. Shinohe, "Smoothing of Si trench sidewall surface by shemical dry etching and sacrificial oxidation," *Jpn. J. Appl. Phys*., vol. 37, p. 3954, 1998.

[21] S. Rauf, P. J. Stout, and J. Cobb, "Modeling the impact of photoresist trim etch process on photoresist surface roughness," *J. Vac. Sci. Technol. B*, vol. 21, pp. 655-659, 2003.

[22] E. Baravelli, A. Dixit, R. Rooyackers, M. Jurczak, N. Speciale, and K. De Meyer, "Impact of line-edge roughness on FinFET matching performance," *IEEE Trans. Electron Devices*, vol. 54, pp. 2466-2474, 2007.

[23] K. Patel, T.-J. King, and C. J. Spanos, "Gate line edge roughness model for estimation of FinFET performance variability," *IEEE Trans. Electron Devices*, vol. 56, pp. 3055-3063, 2009.

[24] S. Yu, Y. Zhao, Y. Song, G. Du, J. Kang, R. Han, and X. Liu, "3-D simulation of geometrical variations impact on nanoscale FinFETs," in *Proc. ICSICT*, 2008, pp. 408-411.

[25] N. Sano, K. Matsuzawa, M. Mukai, N. Nakayama, “On discrete random dopant modeling in drift-difussion simulations: physical meaning of ‘atomistic’ dopants,” *Microelec. Reliability*, vol. 42, pp. 189-199, 2002.

[26] M-H. Chiang, J-N. Lin, K. Kim, and C-T. Chuang, “Random dopant fluctuation in limited-width FinFET technologies,” *IEEE Trans. Electron Devices*, vol. 54, no. 8, pp. 2055-2060, Aug. 2007.

[27] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits*, vol. 24, pp. 1433-1439, 1989.

[28] Predictive Technology Model. [Online]. Available: http://ptm.asu.edu

[29] MIPS [Online]. Available: http://opencores.org

[30] ARM Cortex-M0 processor. [Online]. Available: http://www.arm.com/products/processors/cortex-m/cortex-m0.php

[31] J. A. Power, A. Mathewson, W.A. Lane, “MOSFET statistical parameter extraction using multivariate statistics”, in *Proc. ICMTS*, pp. 209-214, 1991.

[32] K. Takeuchi and M. Hane, "Statistical compact model parameter extraction by direct fitting to variations", *IEEE Trans. Electron Devices*, vol. 55, no. 6, pp. 1487-1493, Jun 2008.

[33] J-P. Colinge, C-W. Lee, A. Afzalian, N. D. Akhavan, R. Yan, I. Ferain, P. Razavi, B. O’Neill, A. Blake, M. White, A-M. Kelleher, B. McCarthy, and R. Murphy, "Nanowire transistors without junctions," *Nature Nanotechnol.*, vol. 5, pp. 225-229, 2010.

[34] S-J. Choi, D-I. Moon, S. Kim, J. P. Duarte, and Y-K. Choi, “Sensitivity of threshold voltage to nanowire width variation in junctionless transistors,” *IEEE Electron Device Lett.*, vol. 32, pp. 125-127, 2011.

[35] A. Kranti, C-W. Lee, I. Ferain, R. Yu, N. D. Akhavan, P. Razavi, J-P. Colinge, “Junctionless nanowire transistor (JNT): Properties and design guidelines,” *Proc. ESSDERC*, 2010, pp. 357-360.

[36] C-W. Lee, I. Ferain, A. Afzalian, R. Yan, N. D. Akhavan, P. Razavi, and J-P. Colinge, “Performance estimation of junctionless multigate transistors,” *Solid-State Electronics*, vol. 54, pp. 97-103, 2010.

[37] S. Gundapaneni, S. Ganguly, and A. Kottantharayil, “Bulk planar junctionless transistor (BPJLT): An attractive device alternative for scaling,” *IEEE Electron Device Lett.*, vol. 32, pp. 261-263, 2011.

[38] R. Rios, A. Cappellani, M. Armstrong, A. Budrevich, H. Gomez, R. Pai, N. Rahhal-orabi, and K. Kuhn, “Comparison of junctionless and conventional trigate transistors with L_g down to 26 nm,” *IEEE Electron Device Lett.*, vol. 32, pp. 1170-1172, Sep. 2011.

[39] R. Trevisoli, R. Doria, M. de Souza, and M. Pavanello, “Threshold voltage in junctionless nanowire transistors,” *Semicond. Sci. Technol.*, vol. 26, 105009, 2011.

[40] M. Aldegunde, A. Martinez, and J. R. Barker, “Study of discrete doping-induced variability in junctionless nanowire MOSFETs using dissipative quantum transport simulations,” *IEEE Electron Device Lett.*, vol. 33, pp. 194-196, 2012.

[41] A. Asenov, “Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 μm MOSFET’s: A 3-D ‘Atomistic’ simulation study,” *IEEE. Trans. Electron Devices*, vol. 45, pp. 2505-2513, 1998.

[42] C. Visweswariah, K. Ravindran, K. Kalafala, S. G. Walker, S. Narayan, D.K. Beece, J. Piaget, N. Venkateswaran, and J. G. Hemmett, "First-order incremental block-based statistical timing analysis," *IEEE Trans. Comput. Aided Design Integr. Circuits Syst.*, vol. 25, no. 10, pp. 2170-2180, Oct. 2006

[43] H. Chang and S. S. Sapatnekar, “Full-chip analysis of leakage power under process variations, including spatial correlations,” in *Proc. 42nd Annu. Design Autom. Conf.*, Jun. 2005, pp. 1–6.

[44] BSIM4 manual [Online]. http://www-device.eecs.berkeley.edu/bsim/

[45] NanGate FreePDK45 Generic Open Cell Library [Online]. http://www.si2.org/openeda.si2.org/projects/nangatelib

[46] B. H. Calhoun, X. L. Yu Cao, K. Mai, L. T. Pileggi, R. A. Rutenbar, and K. L. Shepard, “Digital circuit design challenges and opportunities in the era of nanoscale CMOS,” *IEEE Special Issue Integr. Electron.*, vol. 96, no. 1, pp. 343–365, Feb. 2008.

[47] E. Karl, Y. Wang, Y. Ng, Z. Guo, F. Hamzaoglu, U. Bhattacharya, K. Zhang, K. Mistry, and M. Bohr, “A 4.6GHz 162Mb SRAM design in 22nm Tri-Gate CMOS technology with integrated active VMIN enhancing assist circuitry,” in *Proc. ISSCC*, 2012, pp. 230–232.

[48] B. S. Haran, A. Kumar, L. Adam, and J. Chang, “22 nm technology compatible fully functional 0.1 μm^2 6T-SRAM cell,” in *Proc. IEEE IEDM*, Dec. 2008, pp. 1–4.

[49] A. Singhee and R. A. Rutenbar, “Statistical blockade: Very fast statistical simulation and modeling of rare circuit events and its application to memory design,” *IEEE Trans. Comput. Aided Design Integr. Circuits Syst.*, vol. 28, no. 8, pp. 1176–1189, Aug. 2009.

[50] K. A. Bowman, S. G. Duvall, and J. D. Meindl, “Impact of dieto-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration,” *IEEE J. Solid-State Circuits*, vol. 37, no. 2, pp. 183–190, Feb. 2002.

[51] K. A. Bowman, S. B. Samaan, and N. Z. Hakim, “Maximum clock frequency distribution model with practical VLSI design considerations,” in *Proc. Integr. Circuit Design Technol., Int. Conf.*, 2004, pp. 183–191.

[52] S. M. Burns, M. Ketkar, N. Menezes, K. A. Bowman, J. W. Tschanz, and V. De, "Comparative analysis of conventional and statistical design techniques," in *Proc. ACM/IEEE Design Autom. Conf.*, Jun. 2007, pp. 238–243.

[53] K. K. Bhuwalka, S. Sedlmaier, A. K. Ludsteck, C. Tolksdorf, J. Schulze, and I. Eisele, "Vertical tunnel field-effect transistor," *IEEE Trans. Electron Devices*, vol. 51, no. 2, pp. 279-282, Feb. 2004.

[54] K. Boucart and A. M. Ionescu, "Double-gate tunnel FET with high-k gate dielectric," *IEEE Trans. Electron Devices*, vol. 54, no. 7, pp. 1725-1732, Jul. 2007.

[55] U. E. Avci, R. Rios, K. Kuhn, and I. A. Young, "Comparison of performance, switching energy and process variations for the TFET and MOSFET in logic," *Proc. VLSI Symposium*, 2011, pp. 124-125.

[56] S. H. Kim, S. Agarwal, Z. A. Jacobson, P. Matheu, C. Hu, and T-J. K. Liu, "Tunnel field effect transistor with raised germanium source," *IEEE Electron Device Lett.*, vol. 31, no. 10, pp. 1107-1109, Oct. 2010.

[57] A. S. Verhulst, W. G. Vandenberghe, K. Maex, and G. Groeseneken, "Boosting the on-current of a n-channel nanowire tunnel field-effect transistor by source material optimization," *J. Appl. Physics*, vol. 104, no. 6, pp. 064514-1–064514-10, Sep. 2008.

[58] S. Mookerjea, D. Mohata, R. Krishnan, J. Singh, A. Vallett, A. Ali, T. Mayer, V. Narayanan, D. Schlom, A. Liu, S. Datta, "Experimental demonstration of 100nm channel length $In_{0.53}Ga_{0.47}As$-based vertical inter-band tunnel field effect transistors (TFETs) for ultra low-power logic and SRAM applications," *IEDM Tech. Digest*, 2009, pp. 1-3.

[59] M. Luisier and G. Klimeck, "Performance comparisons of tunneling field-effect transistors made of InSb, carbon, and GaSb-InAs broken gap heterostructures," *IEDM Tech. Digest*, 2009, pp. 1-4.

[60] H. Zhao, Y. Chen, Y. Wang, F. Zhou, F. Xue, and J. Lee, "InGaAs tunneling field-effect-transistors with atomic-layer-deposited gate oxides," *IEEE Trans. Electron Devices*, vol. 58, no. 9, pp. 2990-2995, Sep. 2011.

[61] L. Wang, E. Yu, Y. Taur, and P. Asbeck, "Design of tunneling field-effect transistors based on staggered heterojunctions for ultralow-power applications," *IEEE Electron Device Lett.*, vol. 31, no. 5, pp. 431-433, May 2010.

[62] A. Ionescu and H. Riel, "Tunnel field-effect transistors as energy-efficient electronic switches," *Nature*, vol. 479, no. 7373, pp. 329-337, Nov. 2011.

[63] E. O. Kane, "Theory of tunneling," *J. Appl. Phys.*, vol. 32, no. 1, pp. 83-91, Jan. 1961.

[64] P. Agarwal, M. J. Goossens, V. Zieren, E. Aksen, and J. W. Slotboom, "Impact ionization in thin silicon diodes," *IEEE Electron Device Lett.*, vol. 25, no. 12, pp. 807-809, Dec. 2004.

[65] P. M. Solomon, J. Jopling, D. J. Frank, C. D'Emic, O. Dokumaci, P. Ronsheim, and W. E. Haensch, "Universal tunneling behavior in technologically relevant p/n junction diodes, *J. Appl. Phys*., vol. 95, no. 10, pp. 5800-5812, May 2004.

[66] D. J. Roulston, N. D. Arora, and S. G. Chamberlain, "Modeling and measurement of minority-carrier lifetime versus doping in diffused layers of n+-p silicon diodes," *IEEE Trans. Electron Devices*, vol. ED-29, no. 2, pp. 284-291, Feb. 1982.

[67] A. Schenk, "A model for the field and temperature dependence of Shockley–Read–Hall lifetimes in silicon," *Solid-State Electronics*, vol. 35, no. 11, pp. 1585-1596, May 1992.

[68] G. Paasch and H. Ubensee, "A modified local density approximation: Electron density in inversion layers," *Phys. Stat. Sol. (b)*, vol. 113, no. 1, pp. 165-178, May 1982.

[69] N. Damrongplasit, C. Shin, S. H. Kim, R. A. Vega, and T-J. K. Liu, "Study of random dopant fluctuation effects in germanium-source tunnel FETs," *IEEE Trans. Electron Devices*, vol. 58, no. 10, pp. 3541-3548, Oct. 2011.

[70] F. Conzatti, M. G. Pala, and D. Esseni, "Surface-roughness-induced variability in nanowire InAs tunnel FETs," *IEEE Electron Device Lett*., vol. 33, no. 6, pp. 806-808, Jun. 2012.

[71] K. Boucart, A. M. Ionescu, and W. Riess, "A simulation-based study of sensitivity to parameter fluctuations of silicon tunnel FETs," in *Proc. ESSDERC*, 2010, pp. 345-348.

[72] A. Asenov, A. R. Brown, J. H. Davies, S. Kaya, and G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs," *IEEE Trans. Electron Devices*, vol. 50, no. 9, pp. 1837-1852, Sep. 2003.

[73] J. del Alamo, "Nanometre-scale with III-V compound semiconductors," *Nature*, vol. 479, pp. 317-323, Nov. 2011.

[74] M. Radosavljevic, G. Dewey, J. M. Fastenau, J. Kavalieros, R. Kotlyar, B. Chu-Kung, W. K. Liu, D. Lubyshev, M. Metz, K. Millard, N. Mukherjee, L. Pan, R. Pillarisetty, W. Rachmady, U. Shah, and R. Chau, "Non-planar, multi-gate InGaAs quantum well field effect transistors with high-k gate dielectric and ultra-scaled gate-to-drain/gate-to-source separation for low power logic applications," in *IEDM Tech. Digest*, 2010, pp. 126-129.

[75] J. Lin, D. Antoniadis, and J. del Alamo, "Novel intrinsic and extrinsic engineering for high-performance high-density self-aligned InGaAs MOSFETs: precise channel thickness control and sub-40-nm metal contacts," in *IEDM Tech. Digest*, 2014, pp. 574-577.

[76] J. del Alamo, D. Antoniadis, A. Guo, D-H. Kim, T-W. Kim, J. Lin, W. Lu, A. Vardi, and X. Zhao, "InGaAs MOSFETs for CMOS: Recent advances in process technology," in *IEDM Tech. Digest*, 2013, pp. 24-27.

[77] Y.Q. Wu, R.S. Wang, T. Shen, J.J. Gu and P. D. Ye, "First experimental demonstration of 100 nm inversion-mode InGaAs FinFET through damage-free sidewall etching," in *IEDM Tech. Digest*, 2009, pp. 331-334.

[78] J. J. Gu, X. W. Wang, H. Wu, J. Shao, A. T. Neal, M. J. Manfra, R. G. Gordon, and P. D. Ye, "20-80nm channel length InGaAs gate-all-around nanowire MOSFETs with EOT=1.2nm and lowest SS=63mV/dec," in *IEDM Tech. Digest*, 2012, pp. 633-636.

[79] D.-H. Kim, P. Hundal, A. Papavasiliou, P. Chen, C. King, J. Paniagua, M. Urteaga, B. Brar, Y.G. Kim, J.-M. Kuo, J. Li, P. Pinsukanjana, and Y.C. Kao, "E-mode planar L_g = 35 nm $In_{0.7}Ga_{0.3}As$ MOSFETs with $InP/Al_2O_3/HfO_2$ (EOT = 0.8 nm) composite insulator," in *IEDM Tech. Digest*, 2012, pp. 761-764.

[80] Y. Sun, A. Majumdar, C.-W. Cheng, Y.-H. Kim, U. Rana, R. M. Martin, R. L. Bruce, K.-T. Shiu, Y. Zhu, D. Farmer, M. Hopstaken, E. A. Joseph, J. P. de Souza, M. M. Frank, S.-L. Cheng, M. Kobayashi, E. A. Duch, D. K. Sadana, D.-G. Park, and E. Leobandung, "Self-aligned III-V MOSFETs: Towards a CMOS compatible and manufacturable technology solution," in *IEDM Tech. Digest*, 2013, pp. 48-51.

[81] X. Zhou, Q. Li, C. W. Tang and K. M. Lau, "30nm enhancement-mode $In_{0.53}Ga_{0.47}As$ MOSFETs on Si substrates grown by MOCVD exhibiting high transconductance and low on-resistance," in *IEDM Tech. Digest*, 2013, pp. 48-51.

[82] T. Irisawa, M. Oda, K. Ikeda, Y. Moriyama, E. Mieda, W. Jevasuwan, T. Maeda, O. Ichikawa, T. Osada, M. Hata, Y. Miyamoto and T. Tezuka, "High electron mobility triangular InGaAs-OI nMOSFETs with (111)B side surfaces formed by MOVPE growth on narrow fin structures," in *IEDM Tech. Digest*, 2013, pp. 28-31.

[83] N. Waldron, C. Merckling, L. Teugels, P. Ong, S. A. U. Ibrahim, F. Sebaai, A. Pourghaderi, K. Barla, N. Collaert, and A. V-Y. Thean, "InGaAs gate-all-around nanowire devices on 300mm Si substrates," *IEEE Electron Device Lett.*, vol. 35, no. 11, pp. 1097-1099, Nov. 2014.

[84] S. Natarajan, M. Agostinelli, S. Akbar, M. Bost, A. Bowonder, V. Chikarmane, S. Chouksey, A. Dasgupta, K. Fischer, Q. Fu, T. Ghani, M. Giles, S. Govindaraju, R. Grover, W. Han, D. Hanken, E. Haralson, M. Haran, M. Heckscher, R. Heussner, P. Jain, R. James, R. Jhaveri, I. Jin, H. Kam, E. Karl, C. Kenyon, M. Liu, Y. Luo, R. Mehandru, S. Morarka, L. Neiberg, P. Packan, A. Paliwal, C. Parker, P. Patel, R. Patel, C. Pelto, L. Pipes, P. Plekhanov, M. Prince, S. Rajamani, J. Sandford, B. Sell, S. Sivakumar, P. Smith, B. Song, K. Tone, T. Troeger, J. Wiedemer, M. Yang, and K. Zhang, "A 14nm logic technology featuring 2nd-generation FinFET transistors, air-gapped interconnects, self-aligned double patterning and a 0.0588 μm^2 SRAM cell size," in *IEDM Tech. Digest*, 2014, pp. 71-73.

[85] M. Radosavljevic, G. Dewey, D. Basu, J. Boardman, B. Chu-Kung, J. M. Fastenau, S. Kabehie, J. Kavalieros, V. Le, W. K. Liu, D. Lubyshev, M. Metz, K. Millard, N. Mukherjee, L. Pan, R. Pillarisetty, W. Rachmady, U. Shah, H. W. Then, and R. Chau, "Electrostatics improvement in 3-D tri-gate over ultra-thin body planar InGaAs quantum well field effect transistors with high-k gate dielectric and scaled gate-to-drain/gate-to-source separation," in *IEDM Tech. Digest*, 2011, pp. 765-768.

[86] T. Boykin, G. Klimeck, R. Bowen, and F. Oyafuso, "Diagonal parameter shifts due to nearest-neighbor displacements in empirical tight-binding theory," *Phys. Rev. B*, vol. 66, p. 125207, Sep. 2002.

[87] T. Boykin, G. Klimeck, and F. Oyafuso, "Valence band effective-mass expressions in the sp3d5s* empirical tight-binding model applied to a Si and Ge parametrization," *Phys. Rev. B*, vol. 69, Mar. 2004.

[88] J. M. Ziman, *Principles of the Theory of Solids*. Cambridge: Cambridge University Press, 2nd ed., 1972.

[89] N. Seoane, G. Indalecio, E. Comesaña, A. J. García-Loureiro, M. Aldegunde, and K. Kalna, "Three Dimensional Simulations of Random Dopant and Metal-Gate Workfunction Variability in an $In_{0.53}Ga_{0.47}As$ GAA MOSFET," *IEEE Electron Device Lett*., vol. 34, no. 2, pp. 205-207, Feb. 2013.

[90] N. Seoane, A.J. Garcia-Loureiro, K. Kalna, and A. Asenov, "Random dopant related variability in the 30 nm gate length $In_{0.75}Ga_{0.25}As$ implant free MOSFET," *J. Comput. Electron*., vol. 6, pp. 159-163, Feb. 2008.

[91] NSM Archive: Physical properties of semiconductors [Online]. http://www.ioffe.rssi.ru/SVA/NSM/Semicond/

[92] S. Datta, "Nanoscale device modeling: the Green's function method," *Superlattices and Microstructures*, vol. 28, no. 4, pp. 253-278, Jul. 2000.

[93] T.-H. Yu, E. Hsu, C.-W. Liu, J.-P. Colinge, Y.-M. Sheu, J. Wu, and C. H. Diaz, "Electrostatics and ballistic transport studies in junctionless nanowire transistors," *Proc. SISPAD* 2013, pp. 85–88, Sept. 2013.

[94] P. Razavi and G. Fagas, "Electrical performance of III-V gate-all-around nanowire transistors," *Appl. Phys. Lett*., vol. 103, no. 6, p. 063506, 2013.

[95] G. Roy, A. R. Brown, F. Adamu-Lema, S. Roy, and A. Asenov, "Simulation study of individual and combined sources of intrinsic parameter fluctuations in conventional nano-MOSFETs," *IEEE Trans. Electron Devices*, vol. 53, no. 12, pp. 3063-3070, Dec. 2006.

[96] A. Asenov, F. Adamu-Lema, X. Wang, and S. M. Amoroso, "Problems with the continuous doping TCAD simulations of decananometer CMOS transistors," *IEEE Trans. Electron Devices*, vol. 61, no. 8, pp. 2745-2751, Aug. 2014.

[97] C. Jacoboni, C. Canali, G. Ottaviani, and A. A. Quaranta, "A review of some charge transport properties of silicon," *Solid State Electron*., vol. 20, no. 2, pp. 77-89, Jul. 1977.

[98] A. Gnudi, S. Reggiani, E. Gnani, and G. Baccarani, "Analysis of threshold voltage variability due to random dopant fluctuations in junctionless FETs," *IEEE Electron Device Lett.*, vol. 33, no. 3, pp. 336-338, Mar. 2012.

[99] A. Martinez, A. R. Brown, N. Seoane, and A. Asenov, “Perturbative vs non-perturbative impurity scattering in a narrow Si nanowire GAA transistor: A NEGF study,” *J. Phys.: Conf. Ser*., vol. 193, p. 012047, Nov. 2009.

[100] D. Chattopadhyay and H. Queisser, “Electron scattering by ionized impurities in semiconductors,” *Rev. Mod. Phys*., vol. 53, pp. 745–768, Oct. 1981.

[101] M. Anantram and A. Svizhenko, “Multidimensional Modeling of Nanotransistors,” *IEEE Trans. Electron Devices*, vol. 54, pp. 2100–2115, Sept. 2007.

[102] M. Fischetti, “Effect of the electron-plasmon interaction on the electron mobility in silicon,” *Phys. Rev. B*, vol. 44, pp. 5527–5534, Sept. 1991.

[103] A. Altland, *Condensed matter field theory*. Cambridge, UK ; New York: Cambridge University Press, 2006.

[104] R. Golizadeh-Mojarad and S. Datta, “Nonequilibrium Greens function based models for dephasing in quantum transport,” *Phys. Rev. B*, vol. 75, Feb. 2007.

[105] J. Oh, D. Ahn, Y. Yu, and S. Hwang, “Green-function approach to transport through a gate-surrounded Si nanowire with impurity scattering,” *Phys. Rev. B*, vol. 77, Jan. 2008.

[106] R. Kubo, “Statistical-mechanical theory of irreversible processes. I. General theory and simple applications to magnetic and conduction problems,” *J. Phys. Soc. Jpn*., vol. 12, pp. 570–586, Jun. 1957.

[107] D. A. Greenwood, “The Boltzmann equation in the theory of electrical conduction in metals,” *Proc. Phys. Soc*., vol. 71, pp. 585–596, Apr. 1958.

[108] K.-I. Goto, T.-H. Yu, J. Wu, C. H. Diaz, and J. P. Colinge, “Mobility and screening effect in heavily doped accumulation-mode metaloxide-semiconductor field-effect transistors,” *Appl. Phys. Lett*., vol. 101, p. 073503, Aug. 2012.

[109] T. Rudenko, A. Nazarov, R. Yu, S. Barraud, K. Cherkaoui, P. Razavi, and G. Fagas, “Electron mobility in heavily doped junctionless nanowire SOI MOSFETs,” *Microelec. Eng*., vol. 109, pp. 326–329, Sep. 2013.

[110] M. Diarra, Y.-M. Niquet, C. Delerue, and G. Allan, “Ionization energy of donor and acceptor impurities in semiconductor nanowires: Importance of dielectric confinement,” *Phys. Rev. B*, vol. 75, Jan. 2007.

[111] C. Jacoboni and L. Reggiani, “The Monte Carlo method for the solution of charge transport in semiconductors with applications to covalent materials,” *Rev. Mod. Phys*., vol. 55, pp. 645–705, Jul. 1983.

[112] M. Poljak, V. Jovanovic, D. Grgec, and T. Suligoj, “Assessment of electron mobility in ultrathin-body InGaAs-on-insulator MOSFETs using physics-based modeling,” *IEEE Trans. Electron Devices*, vol. 59, pp. 1636–1643, Jun. 2012.

[113] T. Kubis and P. Vogl, "Assessment of approximations in nonequilibrium Green's function theory," Phys. Rev. B, vol. 83, p. 195304, May 2011.

[114] S. Jin, M. Fischetti, and Ting-Wei Tang, "Modeling of surface roughness scattering in ultrathin-body SOI MOSFETs," *IEEE Trans. Electron Devices*, vol. 54, pp. 2191–2203, Sep. 2007.

[115] A. Martinez, M. Bescond, J. Barker, A. Svizhenko, M. Anantram, C. Millar, and A. Asenov, "A self-consistent full 3-D real-space NEGF simulator for studying nonperturbative effects in nano-MOSFETs," *IEEE Trans. Electron Devices*, vol. 54, pp. 2213–2222, Sep. 2007.

[116] R. Lake, G. Klimeck, R. C. Bowen, and D. Jovanovic, "Single and multiband modeling of quantum electron transport through layered semiconductor devices," *J. Appl. Phys.*, vol. 81, no. 12, pp. 7845–7869, 1997.

[117] D. E. Nikonov, H. S. Pal, and G. Bourianoff, "Scattering in NEGF: Made simple," 2009.

[118] M. Yokoyama, R. Iida, S. Kim, N. Taoka, Y. Urabe, H. Takagi, T. Yasuda, H. Yamada, N. Fukuhara, M. Hata, M. Sugiyama, Y. Nakano, M. Takenaka, and S. Takagi, "Sub-10-nm extremely thin body InGaAs-on-insulator MOSFETs on Si wafers with ultrathin Al_2O_3 buried oxide layers," *IEEE Electron Device Lett*., vol. 32, pp. 1218–1220, Sep. 2011.

[119] P. Osgnach, E. Caruso, D. Lizzit, P. Palestri, D. Esseni, and L. Selmi, "The impact of interface states on the mobility and drive current of $In_{0.53}Ga_{0.47}As$ semiconductor n-MOSFETs," *Solid-State Electron*., vol. 108, pp. 90–96, Jun. 2015.

[120] M. Yokoyama, R. Iida, S. H. Kim, N. Taoka, Y. Urabe, T. Yasuda, H. Takagi, H. Yamada, N. Fukuhara, M. Hata, M. Sugiyama, Y. Nakano, M. Takenaka, and S. Takagi, "Extremely-thin-body InGaAs-on-insulator MOSFETs on Si fabricated by direct wafer bonding," in *IEDM Tech. Dig.* 2010, pp. 1-4, Dec. 2010.

[121] G. Hu, P. Xiang, Z. Ding, R. Liu, L. Wang, and T.-A. Tang, "Analytical models for electric potential, threshold voltage, and subthreshold swing of junctionless surrounding-gate transistors," *IEEE Trans. Electron Devices*, vol. 61, pp. 688–695, Mar. 2014.

[122] A. Pan and C. O. Chui, "Modeling source-drain tunneling in ultimately scaled III-V transistors," submitted, 2015.

[123] M. J. Rosker, V. Greanya, and T-H. Chang, "The DARPA COmpound Semiconductor Materials on Silicon (COSMOS) program," in *Proc. CSIC*, 2008, pp. 1-4.

[124] A. Gutierrez-Aitken, P. Chang-Chien, D. Scott, K. Hennig, E. Kaneshiro, P. Nam, N. Cohen, D. Ching, K. Thai, B. Oyama, J. Zhou, C. Geiger, B. Poust, M. Parlee, R. Sandhu, W. Phan, A. Oki, and R. Kagiwada, "Advanced heterogeneous integration of InP HBT and CMOS Si technologies," in *Microwave Symp. Tech. Digest*, 2009, pp. 1109-1112.

[125] P. Chang-Chien, X. Zeng, K. Tornquist, M. Nishimoto, M. Battung, Y. Chung, J. Yang, D. Farkas, M. Yajima, C. Cheung, K. Luo, D. Eaves, J. Lee, J. Uyeda, D. Duan, O. Fordham, T. Chung, R. Sandhu, and R. Tsai, "MMIC compatible wafer-level packaging technology," in *Proc. IPRM*, 2007, pp. 14-17.

[126] L. Czornomaz, N. Daix, K. Cheng, D. Caimi, C. Rossel, K. Lister, M. Sousa, and J. Fompeyrine, "Co-integration of InGaAs n- and SiGe p-MOSFETs into digital CMOS circuits using hybrid dual-channel ETXOI substrates," in *IEDM Tech. Dig.* 2013, pp. 52-55.

[127] M. Yokoyama, S. Kim, R. Zhang, N. Taoka, Y. Urabe, T. Maeda, H. Takagi, T. Yasuda, H. Yamada, O. Ichikawa, N. Fukuhara, M. Hata, M. Sugiyama, Y. Nakano, M. Takenaka and S. Takagi, "III–V/Ge high mobility channel integration of InGaAs n-channel and Ge p-channel metal–oxide–semiconductor field-effect transistors with self-aligned ni-based metal source/drain using direct wafer bonding," *Appl. Phys. Express*, vol. 5, no. 7, p. 076501, Jun. 2012.

[128] K. L. Lew, S. F. Yoon, W. K. Loke, H. Tanoto, C. L. Dohrman, D. M. Isaacson, and E. A. Fitzgerald. "High gain AlGaAs/GaAs heterojunction bipolar transistor fabricated on SiGe/Si substrate." *J. Vac. Sci. Technol. B*, vol. 25, no. 3 pp. 902-905, May 2007.

[129] Y. Q. Wu, M. Xu, P. D. Ye, Z. Cheng, J. Li, J. Park, J. Hydrick, J. Bai, M. Carroll, J. G. Fiorenza, and A. Lochtefeld, "Atomic-layer-deposited Al_2O_3/GaAs metal-oxide-semiconductor field-effect transistor on Si substrate using aspect ratio trapping technique", *Appl. Phys. Lett.*, vol. 93, pp. 242106-1–242106-3, Dec. 2008.

[130] H. Ko, K. Takei, R. Kapadia, S. Chuang, H. Fang, P. W. Leu, K. Ganapathi, El Plis, H. S. Kim, S-Y. Chen, M. Madsen, A. C. Ford, Y-L. Cheuh, S. Krishna, S. Salahuddin, and A. Javey, "Ultrathin compound semiconductor on insulator layers for high-performance nanoscale transistors," *Nature*, vol. 468, no. 7321, pp. 286-289, Nov. 2010.

[131] K. Takei, R. Kapadia, H. Fang, E. Plis, S. Krishna, and A. Javey, "High quality interfaces of InAs-on-insulator field-effect transistors with ZrO_2 gate dielectrics," *Appl. Phys. Lett.*, vol. 102, no. 15, pp. 153513, 2013.

[132] J. Nah, H. Fang, C. Wang, K. Takei, M. H. Lee, E. Plis, S. Krishna, and A. Javey, "III-V complementary metal-oxide-semiconductor electronics on silicon substrates," *Nano Lett.*, vol. 12, pp. 3592-3595, Jun. 2012.

[133] C. O. Chui, K-S. Shin, J. Kina, K-H. Shih, P. Narayanan, and C. A. Moritz, "Heterogeneous integration of epitaxial nanostructures – strategies and application drivers," in *Proc. SPIE*, 2012, pp. 84670R-1–84670R-15.

[134] J.-H. Ahn, H.-S. Kim, K. J. Lee, S. Jeon, S. J. Kang, Y. Sun, R. G. Nuzzo, and J. A. Rogers," Heterogeneous three-dimensional electronics by use of printed semiconductor nanomaterials," *Science*, vol. 314, no. 5806, pp. 1754-1757, Dec. 2006.

[135] A. Carlson, A. M. Bowen, Y. Huangk, R. G. Nuzzo, and J. A. Rodgers, "Transfer printing techniques for materials assembly and micro/nanodevice fabrication," *Adv. Mater.*, vol. 24, pp. 5284-5318, 2012.

[136] K-H. Shih, "III-V multigate non-planar channel transistor simulations and technologies," Ph.D dissertation, Dept. Elect. Eng., UCLA, Los Angeles, CA 2012.

[137] S. J. Pearton, J. M. Kuo, W. S. Hobson, E. Hailemarian, F. Ren, A. Katz, and A. P. Perley, "Ion implantation doping of InGaP, InGaAs, and InAlAs," *Proc. MRS Symp.*, 1992, pp. 797-804.

[138] K. S. Jones, and E. E. Haller, "Ion implantation of boron in germanium," *J. Appl. Phys.*, vol. 61, no. 7, pp. 2469-2477, 1987

[139] R. T. P. Lee, Y. Oshawa, C. Huffman, Y. Trickett, G. Nakamura, C. Hatem, K.V. Rao, F. Khaja, R. Lin, K. Matthews, K. Dunn, A. Jensen, T. Karpowicz, P. F. Nielsen, E. Stinzianni, A. Cordes, P. Y. Hung, D-H. Kim, R.J.W. Hill, W-Y. Loh, and C. Hobbs, Ultra low contact resistivity ($< 1\times10^{-8}$ Ω-cm^2) to $In_{0.53}Ga_{0.47}As$ fin sidewall (110)/(100) surfaces: Realized with a VLSI processed III-V fin TLM structure fabricated with III-V on Si substrates," *IEDM Tech. Digest*, 2014, pp. 776-779.

[140] J.J. Gu, X.W. Wang, H. Wu, J. Shao, A.T. Neal, M.J. Manfra, R.G. Gordon, and P.D. Ye, "20–80nm channel length InGaAs gate-all-around nanowire MOSFETs with EOT=1.2nm and lowest SS=63mV/dec," *IEDM Tech. Digest*, 2012, pp. 633-636.

[141] I. Ok, D. Veksler, P.Y. Hung, J. Oh, R. L. Moore, C. McDonough, R. E. Geer, C. K. Gaspe, M.B. Santos, G. Wong, P. Kirsch, H. H. Tseng, G. Bersuker, C. Hobbs, and R. Jammy, "Reducing R_{ext} in laser annealed enhancement-mode $In_{0.53}Ga_{0.47}As$ surface channel n-MOSFET," *Proc. VLSI-TSA*, 2010, pp. 38-39.

[142] C. O. Chui, H. Kim, D. Chi, B. B. Triplett, P. C. McIntyre, and K. C. Saraswat, "A sub-400°C germanium MOSFET technology with high-k dielectric and metal gate," *IEDM Tech. Dig.*, pp. 437-440, San Francisco, CA, December 8-11, 2002.

[143] Imprio 450, Molecular Imprints [Online], http://www.molecularimprints.com/products/imprio450.php.

[144] ASML TWINSCAN NXT:1950i Datasheet, ASML.

[145] G. C. DeSalvo, W. F. Tseng, and J. Comas, "Etch rates and selectivities of citric acid/hydrogen peroxide on GaAs, $Al_{0.3}Ga_{0.7}As$, $In_{0.2}Ga_{0.8}As$, $In_{0.53}Ga_{0.47}As$, $In_{0.52}Al_{0.48}As$, and InP," *J. Electrochem. Soc.*, vol. 139, no. 3, pp. 831-835, Mar. 1992.

[146] J-H. Kim, D. H. Lim, and G. M. Yang, "Selective etching of AlGaAs/GaAs structures using the solutions of citric acid/H_2O_2 and de-ionized H_2O/buffered oxide etch," *J. Vac. Sci. Technol. B*, vol. 16, no. 2, pp. 558-560, Mar. 1998.

[147] Y. Sun, V. Kumar, I. Adesida, and J. A. Rogers, "Buckled and wavy ribbons of GaAs for high-performance electronics on elastomeric substrates," *Adv. Mater.*, vol. 18, pp. 2857-2862, 2006.

[148] H. Chen, X. Feng, Y. Huang, Y. Huang, J. A. Rogers, "Experiments and viscoelastic analysis of peel test with patterned strips for applications to transfer printing," *J. Mech. Phys. Solids*, vol. 61, pp. 1737-1752, Apr. 2013.

[149] H.-J. Kim-Lee, A. Carlson, D. S. Grierson, J. A. Rogers, and K. T. Turner, "Interface mechanics of adhesiveless microtransfer printing process," *J. Appl. Phys.*, vol. 115, p. 143513, Apr. 2014.

[150] H. Chen, X. Feng, and Y. Chen, "Directionally controlled transfer printing using micropatterned stamps," *Appl. Phys. Lett.*, vol. 103, p. 151607, Oct. 2013.

[151] M. Arafa, C. Youtsey, R. Grundbacher, I. Adesida, and J. Klem, "Fabrication of nanostructures in AlGaSb/InAs using electron-beam lithography and chemically asisted ion-beam etching," *J. Vac. Sci. Technol. B*, vol. 12, no. 6, pp. 3623-3625, Nov. 1994.

[152] G. C. DeSalvo, R. Kaspi, and C. A. Bozada, "Citric acid etching of $GaAs_{1-x}Sb_x$, $Al_{0.5}Ga_{0.5}Sb$, and InAs for heterostructure device fabrication," *J. Electrochem. Soc.*, vol. 141, no. 12, pp. 3526-3531, Dec. 1994.

[153] B. R. Bennett, "Strain relaxation in InAs/GaSb heterostructures," *Appl. Phys. Lett.*, vol. 73, no. 25, pp. 3736-3738, Dec. 1998.

[154] S. Wang, A. Pan., C. O. Chui, and P. Gupta, "PROCEED: A Pareto optimization-based circuit-level evaluator for emerging devices," *Proc. ASP-DAC*, 2014, pp.818-824.

[155] NanoMOS [Online], https://nanohub.org/resources/nanomos.

[156] A. Rahman, G. Klimeck, and M. Lundstrom, "Novel channel materials for ballistic nanoscale MOSFETs-bandstructure effects," in *Proc. IEDM*, 2005, pp. 604-607.

[157] M. De Michielis, D. Esseni, and F. Druissi, "Analytical models for the insight into the use of alternative channel materials in ballistic nano-MOSFETs," *IEEE Trans. Electron Dev.*, vol. 54, no. 1, pp. 115-123, Jan., 2007.

[158] M. Malloy and L. C. Litt, "Technology review and assessment of nanoimprint lithography for semiconductor and patterned media manufacturing," *J. Micro/Nanolith. MEMS MOEMS*, vol. 10, no. 3, p. 032001, 2011.

[159] B. J. Choi, M. J. Meissl, M. Colburn, T. C. Bailey, P. Ruchhoeft, S. V. Sreenivasan, F. Prins, S. K. Banerjee, J. G. Ekerdt, and C. G. Wilson, "Layer-to-layer alignment for step and flash imprint lithography," in *Proc. SPIE*, vol. 4343, 2001, pp. 436-442.

[160] M. Malloy and L. C. Litt, "Step and flash imprint lithography for semiconductor high volume manufacturing?" in *Proc. SPIE*, vol. 3676, 1999, pp. 379-389.

[161] R. S. Ghaida and P. Gupta, “A framework for early and systematic evaluation of design rules,” in *Proc. ICCAD*, 2009, pp. 615-622.

[162] S. M. Wen and C. O. Chui, “CMOS junctionless FET manufacturing cost evaluation,” *IEEE Trans. Semicond. Manuf*., vol. 26, no. 1, pp. 162-168, Feb. 2013.

[163] P. Banerjee, I. Perez, L. Henn-lecordier, S. B. Lee, and G. W. Rubloff, “Nanotubular metal–insulator–metal capacitor arrays for energy storage,” *Nature Nanotech*., vol. 4, no. 5, pp. 292-296, 2009.

[164] C. C. B. Bufon, J. D. C. González, D. J. Thurmer, D. Grimm, M. Bauer, and O. G. Schmidt, “Self-assembled ultra-compact energy storage elements based on hybrid Nanomembranes,” *Nano Lett*., vol. 10, no. 7, pp. 2506-2510, Jun. 2010.

[165] Z. Liu, Y. Zhan, G. Shi, S. Moldovan, M. Gharbi, L. Song, L. Ma, W. Gao, J. Huang, R. Vajtai, F. Banhart, P. Sharma, J. Lou, P. M. Ajayan, “Anomalous high capacitance in a coaxial single nanowire capacitor, *Nat. Commun.*, vol. *3*, p. 879, Jun. 2012.

[166] A.J. Bard and L.R. Faulkner, *Electrochemical Methods: Fundamentals and Applications*, Wiley, New York, NY, USA 2001.

[167] S. Wang, B. Hsia, C. Carraro, R. Maboudian, “High-performance all solid-state micro-supercapacitor based on patterned photoresist-derived porous carbon electrodes and an ionogel electrolyte,” *J. Mater. Chem. A*. **vol.** *2*, pp. 7997-8002, Feb. 2014.

[168] B. Hsia, J. Marschewski, S. Wang, J. Bin In, C. Carraro, D. Poulikakos, C. P. Grigoropoulos, R. Maboudian, “Highly flexible, all solid-state micro-supercapacitors from vertically aligned carbon nanotubes,” *Nanotechnology*, vol. *25*, p. 055401, Jan. 2014.

[169] M. F. El-Kady, R. B. Kaner, “Scalable fabrication of high-power graphene micro-supercapacitors for flexible and on-chip energy storage,” *Nat. Commun*., vol. 4, p. 1475, Feb. 2014.

[170] D. Pech, M. Brunet, H. Durou, P. Huang, V. Mochalin, Y. Gogotsi, P.L. Taberna, and P. Simon, “Ultrahigh-power micrometre-sized supercapacitors based on onion-like carbon,” *Nature Nanotech.*, vol. 5, no. 9, p. 651-654, Sep. 2010.

[171] L.G. Bettini, M. Galluzzi, A. Podestà, P. Milani, and P. Piseri, “Planar thin film supercapacitor based on cluster-assembled nanostructured carbon and ionic liquid electrolyte,” *Carbon*, vol. 59, p. 212-220, Aug. 2013.

[172] G.-Y. Jung, Z. Li, W. Wu, Y. Chen, D. L. Olynick, S.-Y. Wang, W. M. Tong, and R. S. Williams, “Vapor-phase self-assembled monolayer for improved mold release in nanoimprint lithography,” *Langmuir*, vol. 21, no. 4, pp. 1158-1161, Jan. 2005.

[173] D. Membreno, L. Smith, and B. Dunn, “Silica sol–gel chemistry: creating materials and architectures for energy generation and storage,” *J. Sol-Gel Sci. Tech*. vol. 70, no. 2, pp. 203-215, 2014.

[174] D. Membreno, L. Smith, K.-S. Shin, C. O. Chui, and B. Dunn, "A high-energy-density quasi-solid-state carbon nanotube electrochemical double-layer capacitor with ionogel electrolyte," *Transl. Mater. Res.*, vol. 2, no. 1, p. 015001, 2015.

[175] P. L. Taberna, P. Simon, and J. F. Fauvarque, "Electrochemical characteristics and impedance spectroscopy studies of carbon-carbon supercapacitors," *J. Electrochem. Soc*., vol. 150, no. 3, pp. A292-A300, Jan. 2003.

[176] H. Wang and L. Pilon, "Accurate simulations of electric double-layer capacitance of ultramicroelectrodes," *J. Phys. Chem. C*, vol. 115, pp. 16711-16719, 2011.

[177] K. Shoorideh and C. O. Chui, "On the origin of enhanced sensitivity in nanoscale FET-based biosensors," *Proc. NAS*, vol. 111, 2014, pp. 5111-5116.

[178] J. E. B. Randles, "Kinetics of rapid electrode reactions". *Discuss. Fararaday Soc*., vol. 1, pp. 11-19, Mar. 1947.

[179] J. Leys, M. Wubbenhorst, C. P. Menon, R. Rajesh, J. Thoen, C. Glorieux, P. Nockermann, B. Thijs, K. Binnemans, S. Longuemart, "Temperature dependence of the electrical conductivity of imidazolium ionic liquids," *J. Chem. Phys.*, vol. 128, p. 064509, Feb. 2008.

[180] J. Sangoro, C. Iacob, A. Serghei, S. Naumov, P. Galvosas, J. Karger, C. Wespe, F. Bordusa, A. Stoppa, J. Hunger, R. Buchner, F. Kremer, "Electrical conductivity and translational diffusion in the 1-butyl-3-methylimidazolium tetrafluoroborate ionic liquid," *J. Chem. Phys.*, vol. 128, p. 214509, Jun. 2008.