DEVICE AND ARCHITECTURE CO-DESIGN FOR ULTRA-LOW POWER LOGIC USING EMERGING TUNNELING-BASED DEVICES

A Dissertation in
Computer Science and Engineering
by
Vinay Saripalli

© 2011 Vinay Saripalli

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2011
The dissertation of Vinay Saripalli was reviewed and approved by the following:

Suman Datta  
Professor of Electrical Engineering  
Co-Chair of Committee, Dissertation Co-Advisor

Vijaykrishnan Narayanan  
Professor of Computer Science and Engineering  
Co-Chair of Committee, Dissertation Co-Advisor

Mary Jane Irwin  
A. Robert Noll Chair in Engineering and Evan Pugh Professor of Computer Science and Engineering

Kyusun Choi  
Assistant Professor of Computer Science and Engineering

Jeya Chandra  
Professor of Industrial and Manufacturing Engineering

Raj Acharya  
Professor of Computer Science and Engineering  
Head of the Department, Computer Science and Engineering

*Signatures are on file in the Graduate School.
Abstract

The scaling of silicon CMOS, by delivering lower switching-energy transistors with each technology generation, has been the driving force behind total circuit-energy reduction during the past three decades. However, during the past decade it has become increasingly challenging to achieve energy efficiency through scaling of conventional silicon CMOS. One of the key reasons that has caused this energy scaling challenge is the slowdown in $V_{CC}$ scaling, due to non-scalability of $V_T$. The other reason is the steady increase in leakage power consumption with each technology generation due to worsening short channel effects. As the development of sub-22nm CMOS devices is currently underway, alternative transistor-architectures such as 3D tri-gate, and alternative novel semiconductor material systems such as III-V InGaAs (indium gallium arsenide) are being considered to allow both continued feature size scaling as well as $V_{CC}$ scaling. However, some of these emerging transistors have unique properties which require conventional circuit and system design to be re-examined. In this context, co-design of novel devices and architecture, which is the topic of this dissertation, enables the drive toward continued energy reduction. In this dissertation, circuit and architecture-level design aspects for emerging low-$V_{CC}$ devices, such as the Single Electron Transistor (SET) and the Inter-band Tunneling Field Effect Transistor (TFET), are examined.

The energy-delay characteristics of low-$V_{CC}$ sub-300mV logic circuits based on nearly broken-gap GaSb-InAs (galium antimonide-indium arsenide) heterojunction TFETs, are modeled. By taking advantage of an energy-delay crossover behavior between silicon CMOS and heterojunction TFET logic circuits, a hybrid heterogeneous CMOS-TFET multi-core processor architecture is proposed. Simulated execution of several single and multi-threaded benchmark programs on this heterogeneous architecture shows significant energy-delay advantages compared to either a homogeneous CMOS or TFET multi-core processor alone. Device-level variability plays an important role in restricting the minimum operating voltage
(V_{CC\text{-min}}) at which a circuit can operate in a reliable manner. We have constructed a device-level variability model for TFETs and studied the impact of variation on TFET-based SRAM bit-cells. From this study, it is concluded that a Schmitt Trigger-based TFET SRAM bit-cell with a built-in feedback mechanism is most suitable for enabling ultra-low V_{CC} operation of heterojunction TFET-based SRAM circuits. Following this, the asymmetric sub-threshold characteristics of degenerately-doped source (DDS) n-channel and p-channel III-V Tunnel FETs are examined. A TFET-based 4T-loadless SRAM cell is proposed, taking advantage of the asymmetric sub-threshold characteristics of n and p-channel TFETs, which shows improved performance and lower leakage compared to Si CMOS-based 4T-Loadless SRAM cells at low-V_{CC}. A patent application has been filed based on the TFET-based 4T-loadless SRAM and a provisional patent has been approved. Finally, A design space based on physical dimensions and electrostatic properties is proposed to enable circuit-level analysis for SETs. Subsequent circuit-analysis shows that a Sense Amplifier-Binary Decision Diagram (SA-BDD) circuit-architecture, is most suitable to enable ultra low-V_{CC} sub-150mV for practically realizable Single Electron Transistors, which have low self-gain.

These design studies show that it is necessary to model and study characteristics such as energy-delay performance and variability of emerging transistors, in order to enable ultra low-power circuit operation in the nanoscale regime.
# Table of Contents

**List of Figures**

viii

**List of Tables**

xii

**Acknowledgments**

xiii

**Chapter 1**

1. **Introduction**
   1.1 Motivation ........................................... 1
   1.2 Need For Device-Architecture Co-Design ............... 3
   1.3 The Inter-Band Tunneling Transistor ................ 4
   1.4 The Single Electron Transistor ....................... 6
   1.5 Organization ......................................... 7

**Chapter 2**

An **Energy-Efficient Heterogeneous CMP based on Hybrid TFET-CMOS Cores**

2.1 Introduction ........................................... 8
2.2 Tunnel FET Device Modeling ........................... 10
   2.2.1 Background ......................................... 10
   2.2.2 Device Modelling of III-V homojunction and heterojunction Tunnel FETs ................................ 12
2.3 Characterization of HTFET based Logic and Memory .... 15
   2.3.1 Verilog-A Look-up-Table based TFET Simulation .... 15
   2.3.2 Tunnel FET Complementary Logic .................. 16
   2.3.3 N-Channel Tunnel FET Pass-Transistor Logic ........ 17
   2.3.4 Tunnel FET SRAM Cache ............................ 18
2.4 Architectural Analysis of CMOS and HTFET cores .......................... 20
2.4.1 Processor Abstraction ......................................................... 20
2.4.2 Benchmark Execution .......................................................... 22
2.5 Conclusions ............................................................................. 26

Chapter 3
Variation-Tolerant Ultra Low-Power Heterojunction Tunnel FET SRAM Design 28
3.1 Introduction ................................................................................ 28
3.2 SRAM Cell Design and Characterization ........................................ 29
3.2.1 TFET and CMOS Device Models ............................................. 29
3.2.2 Impact of TFET Late-Saturation Characteristics on SRAM ....... 31
3.2.3 Design of 8T and 10T TFET SRAM .......................................... 33
3.2.3.1 Related Work ............................................................... 33
3.2.3.2 Read Operation in TFET SRAM cells ............................... 33
3.2.3.3 Write Operation in TFET SRAM bit-cell ............................ 35
3.2.3.4 Cell Sizing for TFET ST-1 SRAM cell ............................... 35
3.2.4 Characterization of 8T and 10T TFET SRAM ......................... 38
3.2.4.1 Iso-Area Condition ......................................................... 38
3.2.4.2 Hold and Read Noise Margin Comparison .......................... 38
3.2.4.3 Write Noise Margin Comparison ...................................... 40
3.2.4.4 Energy-Delay Comparison ............................................... 41
3.3 Variation Study ......................................................................... 43
3.3.1 Device-Level Variation Model ............................................... 43
3.3.2 Monte-Carlo Simulation of Read-Write Noise Margins .............. 47
3.4 Conclusion ................................................................................ 50

Chapter 4
Low Power Loadless-4T SRAM cell based on Degenerately Doped Source (DDS) In$_{0.53}$Ga$_{0.47}$As Tunnel FETs 51
4.1 Introduction ................................................................................ 51
4.2 Source- Doping Dependant Sub- Threshold Slope Characteristics .... 51
4.3 A 4T-Loadless SRAM Cell Based on DDS III-V TFETs ................. 53
4.4 Conclusion ................................................................................ 55

Chapter 5
Energy-Delay Performance of Nanoscale Transistors Exhibiting Single Electron Behavior and Associated Logic Circuits 57
5.1 Introduction ................................................................................ 57
5.2 Analytical Modeling of Single Electron Transistors ..................... 58
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2.1 Device Electrostatics</td>
<td>58</td>
</tr>
<tr>
<td>5.2.2 Analytical Expressions for Peak and Valley Currents</td>
<td>60</td>
</tr>
<tr>
<td>5.2.3 Modeling Quantization Effect</td>
<td>63</td>
</tr>
<tr>
<td>5.3 SET Design Space Exploration</td>
<td>66</td>
</tr>
<tr>
<td>5.3.1 Description of SET Design Space</td>
<td>66</td>
</tr>
<tr>
<td>5.3.2 SET- Device Design Case Study</td>
<td>68</td>
</tr>
<tr>
<td>5.4 Circuit Design Using SETs</td>
<td>71</td>
</tr>
<tr>
<td>5.4.1 Logic Gate Operation using SETs</td>
<td>71</td>
</tr>
<tr>
<td>5.4.2 Sense Amplifier-Based Design using SETs</td>
<td>73</td>
</tr>
<tr>
<td>5.4.3 SRAM cells using SETs</td>
<td>76</td>
</tr>
<tr>
<td>5.5 Conclusion</td>
<td>78</td>
</tr>
</tbody>
</table>

**Chapter 6**

**Conclusion**

6.1 Future Work                              80

**Appendix A**

Derivation of Analytical Expressions for $I_{\text{Valley}}$ and $I_{\text{Peak}}$ in a Single Electron Transistor 81

Bibliography                                84
## List of Figures

1.1 Non-scalability of $V_T$. Slowdown of $V_{CC}$ scaling .......................... 1  
1.2 Progression of Moore’s law: past and future technology nodes .... 2  
1.3 III-V FETs as a replacement for strained-Si MOSFET ................. 3  
1.4 Schematics of thin-body n-MOSFET and n-TFET. Band-diagram illustration of sub-threshold operation of n-MOSFET and n-TFET. 4  
1.5 Band-diagram illustration of on-state in n-MOSFET and n-TFET. $I_D$-$V_G$ of n-MOSFET and n-TFET ............................. 5  
1.6 Single Electron Transistor (SET) device schematic .................... 6  

2.1 Experimental, simulated and projected $I_D$-$V_G$ characteristics of $In_{0.53}Ga_{0.47}As$ homojunction TFET ................................. 11  
2.2 Transmission Electron Microscopy (TEM) image of a vertical-mesa $In_{0.53}Ga_{0.47}As$ homojunction TFET ................................. 12  
2.3 Schematic of a vertically-oriented Ultra Thin Body (UTB) TFET structure ................................................................. 12  
2.4 Schematic and band-diagram comparison of heterojunction and homojunction TFETs ................................................. 13  
2.5 Calibration of TCAD simulations with atomistic simulations .... 13  
2.6 $I_{on}$ versus $I_{on}$/$I_{off}$ comparison between MOSFET, heterojunction TFET and homojunction TFET ................................. 14  
2.7 Verilog-A small-signal model for TFET simulation .................... 15  
2.8 Heterojunction NTFET and PTFET - device schematic and operation 16  
2.9 Energy-Frequency performance comparison of CMOS and HTFET complementary digital logic ................................. 17  
2.10 TFET source-drain asymmetry and resulting asymmetric $I_D$-$V_D$ characteristics .................................................. 18  
2.11 Validation of Horowitz approximation for TFETs .................... 19  
2.12 Energy-delay performance comparison and leakage power comparison for CMOS and HTFET-based L1 SRAM cache .................... 20  
2.13 Using a ring-oscillator to model the frequency and power of an Intel Atom processor ............................................. 21
2.14 Voltage-frequency operating points for benchmark execution on HTFET and CMOS processors ................................ 22
2.15 Summary of processor frequency distribution, normalized delay and normalized EDP for energy-aware DVFS policy on benchmarks ... 23
2.16 Summary of normalized delay, normalized energy, normalized EDP and normalized ED² for IPC-Aware DVFS on benchmarks .... 24
2.17 L2-cache miss rates for various benchmarks ................................. 25
2.18 Illustration of multi-core execution for iso-peformance to CMOS @ 1375 MHz ....................................................... 26

3.1 Calibration of DG-MOSFET TCAD simulation with experimental Lg 80nm Si FinFET .................................................. 30
3.2 Structurally identical UTB MOSFET and TFET devices for studying variability impact .............................................. 30
3.3 Comparison of Ids-Vds characteristics for UTB HTFET and UTB Si DG-MOSFET .......................................................... 31
3.4 Voltage Transfer Characteristic (VTC) comparison for HTFET and Si CMOS ................................................................. 31
3.5 TFET and CMOS 6T SRAM cell schematic and static-read noise margin comparison .................................................. 32
3.6 Read-operation in various 8T and 10T TFET SRAM cell configurations .......................................................... 34
3.7 Write-operation in various 8T and 10T TFET SRAM cell configurations .......................................................... 36
3.8 Static Noise Margins vs. Pull-up Ratio (PR) for TFET ST-1 SRAM Cell and Static Noise Margins as a function of Vcc for a Pull-up Ratio of 0.15 .................................................. 37
3.9 Comparison of Hold-Static Noise Margin of various SRAM cells . 39
3.10 Comparison of Read-Static Noise Margin of various SRAM cells . 39
3.11 Write-Noise Margin for various SRAM cells ................................. 40
3.12 Access-delay comparison for various SRAM cells ................. 41
3.13 Read-Access energy comparison for various SRAM cells .... 41
3.14 Illustration of UTB HTFET variation model ..................... 43
3.15 Illustration of Gate-Source overlap in an UTB heterojunction TFET 44
3.16 Change in effective band-gap with Tch due to structural quantization 44
3.17 Small-signal variation coefficients for NMOS, Comparison of statistical distribution of on-current obtained through Monte-Carlo vs. Analytical for NMOS, Small-signal variation coefficients for HTFET, Comparison of statistical distribution of on-current obtained through Monte-Carlo vs. Analytical for HTFET .............. 46
3.18 Components of on-current variance in silicon NMOS and HTFET due to various sources of variation @ different $V_{CC}$

3.19 Comparison of $\% \sigma I_{On}/I_{On}$ for silicon NMOS and HTFET @ different $V_{CC}$

3.20 Probability of read-upset for various CMOS and TFET SRAM bit-cell configurations

3.21 Dynamic Energy and Leakage Power for different SRAM bit-cells operating at their $V_{CC}$-min normalized to TFET ST-2 cell

4.1 Asymmetric sub-threshold slope characteristics of DDS NTFET and PTFET

4.2 Energy band-diagram views of DDS In$_{0.53}$Ga$_{0.47}$As NTFET and PTFET illustrating source degeneracy and minority carrier populations

4.3 Operation of 4T-Loadless SRAM cell

4.4 Comparison of read-access delay for TFET and CMOS Loadless-4T SRAM cells

4.5 Write-upset delay comparison in CMOS and TFET Loadless-4T SRAM cells

5.1 Electrostatic potential of the quantum-dot at $I_{Valley}$ as a function of total self capacitance and drain bias

5.2 Description of single electron tunneling as a Stochastic Markov chain process and as a Birth-Death process

5.3 Contributions of various tunneling rates to SET Coulomb oscillation peaks

5.4 Comparison of $I_{Peak}$ and $I_{Valley}$ obtained through Monte-Carlo simulation and analytical calculation

5.5 Comparison of self charging energy and quantization energy of silicon quantum-dot

5.6 Comparison of $I_{Peak}$ and $I_{Valley}$ obtained through analytical calculation and Monte-Carlo simulation, including quantization energy ($E_Q$) term

5.7 Description of SET device design space

5.8 Switching Slope ($\Delta V_G/\Delta \log(I)$) contour plots as a function of quantum-dot dimension and applied drain bias for different gate-control ratios

5.9 Structural definition of a programmable SET device with split-gate tunable tunnel barriers

5.10 Functional simulation of SiGe/strained-silicon split-gate SET

5.11 Monte-Carlo simulation of SiGe/strained-silicon split-gate SET
5.12 Scanning Electron Micrograph (SEM) image of a programmable SET device with split-gate tunable tunnel barriers . 71
5.13 Measured electrical characteristics of programmable SET showing three different modes of operation . 71
5.14 $I_D-V_G$ and $I_D-V_D$ simulations of n-type and p-type Single Electron Transistors . 71
5.15 Voltage Transfer Characteristics (VTC) of SET inverters with strong and weak gate-control . 72
5.16 2-input nand ring oscillator Energy-Delay performance for different activity factors . 73
5.17 8-input Xor logic using BDD-Sense Amplifier . 74
5.18 8-input Xor Logic using complementary gates . 74
5.19 Energy-Delay performance of 8-input Xor logic using Complementary Gates and BDD-Sense Amplifier logic . 75
5.20 Schematic of Negative Differential Conductance (NDC) block . 76
5.21 Room temperature MC simulation of output characteristics of NDC block with increasing gate control . 76
5.22 Schematic of SRAM based on NDC blocks . 77
5.23 Load-line plot for single-electron SRAM showing the operating reference VCC and the stable binary states . 77
5.24 Layout comparison of 6T CMOS SRAM and 3T hybrid-NDC SRAM 77
5.25 Area, Power, Delay Figure-of-Merit comparison for 6T CMOS SRAM and 3T hybrid MuQFET SRAM . 77
List of Tables

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>TCAD simulation parameters for n-channel heterojunction TFET</td>
<td>14</td>
</tr>
<tr>
<td>2.2</td>
<td>Simics simulation parameters.</td>
<td>22</td>
</tr>
<tr>
<td>3.1</td>
<td>Scaled Si DG-MOSFET simulation parameters.</td>
<td>30</td>
</tr>
<tr>
<td>3.2</td>
<td>Summary of SRAM cell sizing for iso-area comparison (W is the</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>nominal width of a transistor).</td>
<td></td>
</tr>
<tr>
<td>3.3</td>
<td>Sources of Variation for Ultra-Thin-Body Device.</td>
<td>45</td>
</tr>
</tbody>
</table>
Acknowledgments

I would like to express my deepest sense of gratitude to my advisors, Dr. Suman Datta and Dr. Vijaykrishnan Narayanan, for giving me an opportunity to work in the exciting area of device-circuit co-design based on novel semiconductor devices. I have undergone significant intellectual and personal growth under their supervision. Dr. Datta’s novel ideas on semiconductor devices and Dr. Vijaykrishnan’s keen insight into circuit and system-level design, and their mutual co-operation has allowed me to carry out this interdisciplinary research. I owe much of this work to them.

I would also like to express my gratitude to my colleagues Dheeraj Mohata, Ashkar Ali, Ashish Agrawal and Lu Liu for many insightful discussions on semiconductor device physics, device fabrication and circuit design. I would like to convey my special thanks and appreciation to the helpdesk staff and the administrative staff at the Department of Computer Science and Engineering for their prompt help and cooperation during my studies at Penn State.
Dedication

I would like to dedicate this dissertation to my parents, Ramana and Lakshmi, for their unconditional love, support and understanding during the course of my studies.
Chapter 1

Introduction

1.1 Motivation

The vision for scaling electronic components on an integrated circuit was first articulated by G. Moore [1], following which a set of rules for reliable scaling of silicon CMOS devices was formulated by R. Dennard [2]. Dennard’s constant-electric field scaling law requires geometric scaling, as well as voltage scaling, by a constant factor (k), in order to preserve the transistor reliability and power density across technology generations. However, as shown in Fig. 1.1, it has become clear that the scaling of $V_T$ with a constant scaling factor cannot be continued because of sub-threshold leakage and circuit noise-immunity limitations - $V_T$ is now maintained at $\sim 0.2V-0.3V$ in modern technology generations[3]. Non-scalability of $V_T$ had major consequences: (i) supply voltage scaling slowed down considerably because a sufficient overdrive ($V_{CC}/V_T$) is required to satisfy performance demands, thus leading to a power reduction challenge in modern CMOS technologies and (ii) physical dimension scaling without $V_{CC}$ reduction resulted in increased net elec-
Fig. 1.2. Progression of Moore’s Law: Past and future technology nodes.

...tric field strength ($E_{\text{eff}}$), which not only caused increased off-state leakage, but also caused an additional performance hit, because effective-mobility degradation results from increased $E_{\text{eff}}$ [4].

Fig. 1.2 shows the evolution of semiconductor transistors for past and future logic technology nodes. Despite the non-scalability of $V_T$, Moore’s law has progressed in an astounding manner during the last decade (2000-2009). Innovations such as uniaxial tensile and compressive strain for NMOS and PMOS respectively [5], High-$\kappa$/TiN gate-stack on silicon [6], and raised source-drain regions [7] acted as additional performance boosters, which allowed reduced $V_{\text{CC}}$ operation while keeping $V_T$ fixed. As the development of sub-22nm CMOS devices is currently underway, the following trends have emerged to allow both continued feature size scaling, as well as $V_{\text{CC}}$ scaling: (i) the need for improved electrostatics is forcing the semiconductor industry to consider 3D FinFET device-architectures [8, 9], and (ii) the need for low-$V_{\text{CC}}$ operation is forcing the consideration of low-bandgap, high-mobility III-V semiconductor materials [10, 11]. For sub-15nm, sub-300mV
logic technologies, Interband Tunneling Transistors (TFETs) [12, 13, 14, 15, 16] and Single Electron Transistors (SETs) [17, 18, 19, 20, 21], which are introduced in sections 1.3 & 1.4, are some of the novel devices which are being considered by the research community for the post-silicon CMOS era.

1.2 Need For Device-Architecture Co-Design

As shown in Fig. 1.3, III-V MOSFETs [22, 23] are intended as a drop-in replacement for strained-Si MOSFETs. III-V MOSFETs are being engineered to provide equivalent performance to state-of-the-art Si MOSFETs while operating at 0.5V. The lower carrier transport mass in III-V materials allows sufficiently high carrier velocity, so that even at low electric fields and low sheet charge densities, better drive currents are produced than strained silicon CMOS at 0.5V [23]. TFETs and SETs are meant for low-\(V_{CC}\) operation (Fig. 1.2), though they are not drop in replacements because being tunneling-based devices, they cannot deliver high on-currents like MOSFETs. Circuit and system design must be re-examined to understand how more energy-efficient operation than state-of-the-art CMOS can be achieved using these devices. The research in this dissertation tries to address whether scaled versions of these devices will be able to achieve more energy-efficient logic operation than state-of-the-art Si CMOS and what special structures need to be used in order to enable ultra low-\(V_{CC}\) operation. In this context, the focus of this dissertation is co-design of novel devices and architecture in order to enable the drive towards continued energy reduction.
1.3 The Inter-Band Tunneling Transistor

Figs. 1.4A & 1.4B compare the schematics of a thin-body n-channel MOSFET and n-channel TFET. The TFET is a device with structural asymmetry having a P++/i/N+ structure, as opposed to the MOSFET which has a symmetric N++/i/N++ structure. In the MOSFET, one of the heavily N-doped (N++) regions acts as the source, and the other N++ region acts as the drain. Fig. 1.4C shows that the off-current of the n-MOSFET is caused by thermionic emission of carriers from the tail of the Fermi-Dirac distribution over the P-N junction barrier. The temperature dependence of this carrier population in the source region causes the sub-threshold slope of the MOSFET to have a $kT/q$ dependence, with a lower limit of $kT/q\ln10$. This sub-threshold conduction mechanism in a n-MOSFET results in a 60 mV/decade sub-threshold slope (lower limit) at room temperature, as shown in Fig 1.5(C). In the TFET, the heavily P-doped region (P++) acts as the source, and the moderately N-doped region (N+) acts as the drain. Fig. 1.4D shows that in the off-state the tail of the Fermi-Dirac distribution of carriers is filtered by the energy band-gap. Due to this energy-filtering effect, the TFET can transition from a very low leakage state to a high on-current state within a
small $V_G$ step, resulting in a sub-60 mV/decade sub-threshold slope, as shown in Fig. 1.5D.

Further application of gate voltage in the n-MOSFET causes channel inversion resulting in high on-current, as shown in Fig. 1.5C. The on-state conduction in the n-MOSFET is due to injection of electrons from the source, whose velocity profile in a silicon device is shown in Fig. 1.5A. In the n-TFET, further application of gate voltage results in high on-current due to band-to-band generation, as shown in Fig. 1.5D. In the on-state of the n-TFET, electron and hole pairs are generated through band-to-band tunneling across the reverse biased P-N junction barrier, whose generation rate in an InGaAs device is shown in Fig. 1.5B. Figs. 1.5A & 1.5B illustrate that the n-MOSFET and the n-TFET differ significantly in their on-state conduction mechanism.

A wealth of tools and literature is available on simulation of nanoscale MOS-FETs [24, 25, 26, 27, 28] and TFETs [29, 30, 31, 32]. For MOSFET simulations, we use PTM simulation models [25], as well as TCAD simulations [33]. TCAD simulations for MOSFETs are performed using a modified drift-diffusion field-dependent mobility model, as discussed in [34]. Tunnel FET simulations are performed using
1.4 The Single Electron Transistor

Classically, a Single Electron Transistor (SET), shown in Fig. 1.6A, is a Coulomb Blockade device [35] which exhibits Peaks and Valleys in the $I_D-V_G$, as shown in Fig. 1.6B, due to the influence of the self-charging Energy $E_0$ of the device (Given by $E_0 = q^2/C_{\Sigma}$, where $C_{\Sigma}$ is the self-capacitance). In order to realize these Coulomb oscillations, it is necessary for the self charging energy $E_0$ to be dominant over thermal fluctuations ($E_0 >> k.T$, $k$ is Boltzmann constant and $T$ is temperature), and it is also required that the tunneling resistance $R_T$ be significantly larger than the quantum of resistance ($R_T >> h/q^2$, $h$ is Planck’s constant and $q$ is the charge of an electron), in order for quantum fluctuations to be minimal ($h$ is Planck’s constant and $q$ is the charge of the electron). These two conditions are necessary for charge to be localized on the quantum-nanodot when it is in Coulomb blockade mode (i.e. the Off state). In order to build switching logic circuits using SETs, it is necessary to operate the device within the region of transconductance occurring from $I_{\text{Valley}}$ to $I_{\text{Peak}}$ (Fig. 1.6B). SET simulations are done using SIMON, a Monte-Carlo simulation tool [36, 35].
1.5 Organization

This dissertation is organized into 6 chapters. Chapter 2 presents a study on enabling low-$V_{CC}$ sub-300mV logic operation using a heterogeneous CMOS-TFET multi-core architecture. Chapter 3 presents a study on low-$V_{CC}$ variation tolerant TFET SRAM bit-cells. Chapter 4 presents an analysis of p-channel TFETs. Chapter 5 presents analytical modeling and energy-delay analysis of Single Electron Transistor-based circuits for sub-150mV logic and SRAM operation. Chapter 6 concludes the thesis.
Chapter 2

An Energy-Efficient Heterogeneous CMP based on Hybrid TFET-CMOS Cores

2.1 Introduction

Power consumption is a critical constraint hampering progress towards more sophisticated and powerful processors. A key challenge to reducing power consumption has been in reducing the supply voltage due to concerns of either reducing performance (due to reduced drive currents) or increasing leakage (when reducing threshold voltage simultaneously). The sub-threshold slope of the transistor is a key factor in influencing the leakage power consumption. With a steep sub-threshold device it is possible to obtain high drive currents ($I_{on}$) at lower voltages without increasing the off-state leakage current ($I_{off}$). In this chapter, we propose the use of Inter-band Tunneling Field Effect Transistors (TFETs) that can exhibit sub-threshold slopes steeper than the theoretical limit of 60 mV/decade found in MOSFETs. Consequently, TFETs can provide higher performance than MOSFET-based designs at lower supply voltages. However, at higher voltages, the $I_{on}$ of MOSFETs are much larger than can be accomplished by the tunneling mechanism employed in existing TFET devices. This trade-off enables architectural innovations through use of heterogeneous systems that employ both TFET
and CMOS based circuit elements.

Heterogeneous chip-multiprocessors that incorporate cores with different frequencies, micro-architectural resources, instruction-set architectures [37, 38] are already emerging. In all these works, the energy-performance optimizations are performed by appropriately mapping the application to a preferred core. In this work, we add a new technology dimensionality to this heterogeneity by using a mix of TFET and CMOS based cores. The feasibility of TFET cores is analyzed by showing design and circuit simulations of logic and memory components that utilize TFET based device structure characterizations.

Dynamic voltage and frequency scaling (DVFS) is widely used to reduce processor power consumption. The proposed heterogeneous architecture enables to extend the range of operating voltages that are supported, using TFET cores that are efficient at low voltages and CMOS cores that are efficient at high voltages. For an application that is constrained by factors such as I/O or memory latencies, low voltage operation is possible, sacrificing little performance. In such cases a TFET core may be preferable. However, for compute intensive performance critical applications, MOSFETs operating at higher voltages are necessary. This study using two DVFS schemes shows that the choice of TFET or CMOS for executing an application varies based on the intrinsic characteristics of the applications. In a multi-programmed environment which is common on platforms ranging from cell-phones to high-performance processors, the proposed heterogeneous architectures can improve energy efficiencies by matching the varied characteristics of different applications.

The emerging multi-threaded workloads provide an additional dimension to this TFET-CMOS choice. Multi-threaded applications with good performance scalability can achieve much better energy efficiencies utilizing multiple cores operating at lower voltages. While energy efficiency through parallelism is not a new idea, the choice of TFET versus CMOS for the application will change based on the actual voltage at which the cores operate and the degree of parallelism (number of cores). This exploratory study shows that TFET-based cores become more preferred in multi-threaded applications from both energy and performance perspective.
The rest of this chapter is organized as follows:

1. In section 2.2, we discuss in detail how the Tunnel FET is modeled using the device simulator TCAD Sentaurus [33]. We also discuss III-V semiconductor-based homojunction and heterojunction TFETs, and show the advantage in using the latter over the former. By comparing the transistor characteristics of TFETs with state-of-the-art MOSFETs, we identify the potential impact of III-V semiconductor-based heterojunction TFETs at the architecture level.

2. In section 2.3, we demonstrate circuit modeling using the TFET device models from section 2.2, and compare the energy-delay performance of logic and memory elements for MOSFETs and heterojunction TFETs.

3. In section 2.4, we illustrate an abstraction to estimate processor-level energy-delay performance for individual CMOS and TFET processor cores using the circuit modeling techniques developed in section 2.3. By simulating various single and multi-threaded benchmark programs, we show the energy-delay benefits of a heterogeneous CMP architecture with TFET-CMOS processor cores.

4. In section 2.5, we summarize key conclusions from the research presented in this chapter.

2.2 Tunnel FET Device Modeling

2.2.1 Background

Interband Tunnel Field Effect Transistors (TFETs) have garnered tremendous interest in recent years because they show a characteristic sub-60 mV/decade sub-threshold slope (SS) behavior as opposed to conventional CMOS, which are limited to a 60 mV/decade SS. Low-V_{CC} (< 300mV) operation of CMOS is challenging because of the availability of a limited overdrive (V_G – V_T). Allowing stronger drive currents at low-V_{CC} for CMOS would require reduction of V_T which in-turn would result in unacceptable increase in off-state leakage - This is a fundamental limitation in CMOS due to the 60 mV/dec sub-threshold slope limitation. In
contrast, Interband TFETs can deliver sufficiently high drive currents at low $V_{CC}$ while maintaining a good $I_{On}/I_{Off}$ ratio, because they are not limited by a 60 mV/decade sub-threshold slope.

A number of TFETs have been experimentally demonstrated in recent literature, showcasing the progress in fabrication and experimental demonstration of novel tunneling devices. The first such demonstration was an In$_{0.53}$Ga$_{0.47}$As homojunction Tunnel FET [12] illustrating the concept of a vertical interband tunneling transistor. A vertically-oriented, gate-all-around silicon nanowire was demonstrated recently, showing 50 mV/decade over 3 decades of drain current, thus experimentally illustrating a sub-60 mV/decade SS [13]. A horizontally-oriented Ultra-Thin-Body (UTB) InAs-on-silicon TFET was also demonstrated recently showing the utility of a III-V semiconductor layer-transfer-technique in TFET fabrication [14]. A process flow for the creation of a side-gated vertical-mesa TFET has also been demonstrated [15, 16]. Thus, there have been continuous improvements in the experimental demonstration of various tunneling structures, bringing closer the promise of sub-60 mV/decade SS operation.
Fig. 2.2. Transmission Electron Microscopy (TEM) image of a vertical-mesa In$_{0.53}$Ga$_{0.47}$As homojunction TFET [15]

Fig. 2.3. Schematic of a vertically-oriented Ultra Thin Body (UTB) interband TFET structure

2.2.2 Device Modelling of III-V homojunction and heterojunction Tunnel FETs

We use the device simulator TCAD Sentaurus [33] in order to model the $I_D$-$V_G$ characteristics of TFETs. TCAD simulations for TFETs can be calibrated such that they match the characteristics of reported experiments. Fig. 2.1A compares the experimental and simulated characteristics for a single-gate homojunction In$_{0.53}$Ga$_{0.47}$As TFET from [12], and shows a good match between experimental and simulated curves. The parameters used for simulating the single-gate homojunction TFET are from [12]. Using a High-$\kappa$ gate oxide ($\epsilon_{ox}$ 21, $T_{ox}$ 2.5nm), and a double-gated Ultra-Thin-Body (UTB) TFET structure, we obtain projected characteristics of a homojunction TFET as shown in Fig. 2.1B. Thus, it is clear from Fig. 2.1 that a highly scaled UTB TFET structure with a scaled gate-oxide, is necessary for improving TFET on-current performance as well as the sub-threshold slope.

A process flow for fabricating a $0.25 \times 5 \, \mu m^2$ vertical-mesa In$_{0.53}$Ga$_{0.47}$As homojunction TFET with side-gates, whose TEM is shown in Fig. 2.2, was developed in [15]. This process flow was developed with the intention of allowing mesa-scaling in order to achieve a highly scaled vertically-oriented Ultra-Thin-Body (UTB) double-gated TFET structure, whose 3D schematic is shown in Fig. 2.3. Further, this process flow was also extended to a heterojunction TFET (HTFET) in [16]. A schematic and band diagram comparison of an UTB GaSb-InAs heterojunction
TFET and an UTB In$_{0.53}$Ga$_{0.47}$As homojunction TFET is shown in Fig. 2.4. Such UTB TFETs have also been modeled using advanced atomistic simulations (with full band-structure calculations) for the homojunction case [39], as well as the heterojunction case [31]. The TCAD Sentaurus simulations for TFETs in this dissertation are calibrated such that they produce the same transfer characteristics as the atomistic simulations, as shown in Fig. 2.5. The TCAD Sentaurus parameters used for simulating the HTFETs are shown in Table 2.1.

By using the HTFET, a higher $I_{On}$ can be obtained because the staggered P-
N heterojunction provides a higher critical-field strength for efficient inter-band tunneling. In order to understand the circuit level implications of using HTFETs, we compare the $I_{On}$ versus $I_{On}/I_{Off}$ characteristics for the transistor candidates by considering different operating points along the $I_D$-$V_G$ curve for a given $V_{CC}$ window, as shown in Fig. 2.6. Fig. 2.6A shows that at $V_{CC}$ 0.8V, the highest $I_{On}$ and $I_{On}/I_{Off}$ ratio are provided by 22nm CMOS, making it the preferred device for operation at high $V_{CC}$. However, at $V_{CC}$ 0.3V, the CMOS device cannot provide both a good $I_{On}$ as well as a good $I_{On}/I_{Off}$ ratio because of the 60 mV/decade limit on the sub-threshold slope. The $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ homojunction TFET can provide a

![Fig. 2.6](image)

**Fig. 2.6.** Comparison of $I_{On}$ versus $I_{On}/I_{Off}$ ratio for different operating points on the $I_D$-$V_G$ for (A) a $V_{CC}$ window of 0.8V and (B) a $V_{CC}$ window of 0.3V.
good $I_{on}/I_{off}$ but cannot provide a high $I_{on}$ since the homojunction does not allow a strong tunneling current. In contrast, the heterojunction TFET can provide a good $I_{on}$ (due to the staggered P-N junction), as well as a good $I_{on}/I_{off}$, due to the sub-60 mV/decade sub-threshold slope, making it the preferred device for operation at low $V_{CC}$.

2.3 Characterization of HTFET based Logic and Memory

2.3.1 Verilog-A Look-up-Table based TFET Simulation

We capture the transfer characteristics of the TFET obtained through device simulation across a range of voltages in a Verilog-A lookup table, in order to perform circuit simulations. The $I_{DS}(V_{GS},V_{DS})$, $C_{GD}(V_{GS},V_{DS})$ and the $C_{GS}(V_{GS},V_{DS})$ characteristics are captured in two-dimensional look-up tables for modeling TFETs. Fig. 2.7A shows the Verilog-A small-signal model for TFETs, which uses the look-up tables for circuit simulation. Fig. 2.7B and 2.7C show the Voltage Transfer
Characteristics (VTC) and the transient output characteristic of a In$_{0.53}$Ga$_{0.47}$As homojunction TFET inverter ($V_{CC}$ 0.5V), which shows the validity of the Verilog-A look-up-table based method.

### 2.3.2 Tunnel FET Complementary Logic

We analyze the energy-performance characteristics of logic gates constructed using CMOS transistors and HTFETs. We use a predictive BSIM model [25] for 22nm CMOS ($V_T$ 0.2V) which provides an $I_{On}$ of 1.4 mA/um and an $I_{On}/I_{Off}$ of $3 \times 10^3$ when operating at its nominal $V_{CC}$ of 0.8V. We also use a GaSb/InAs HTFET which provides an $I_{On}$ of 100 uA/um and an $I_{On}/I_{Off}$ of $2 \times 10^5$ at $V_{CC}$ 0.3V.

In order to build logic gates, a pull-up device is also required. A p-HTFET can be constructed using a heterojunction with InAs as the source, as shown in Fig. 2.8B. When positive gate and drain voltages are applied to the n-HTFET (Fig. 2.8A), electrons tunnel from the GaSb source into the InAs channel (Fig. 2.8B). In contrast, when negative gate and drain voltages are applied to the p-HTFET (Fig. 2.8C), holes tunnel into the GaSb channel from the InAs source. The n-HTFET and p-HTFET illustrated in Fig. 2.8 are simulated separately using TCAD sentaurus models. The work function used for n-HTFET is 4.85 eV and for p-HTFET is 4.69 eV. A detailed analysis of p-channel TFET characteristics is presented in chapter 4.

We obtain the energy-delay characteristics of HTFET logic gates using the

![Fig. 2.8.](image)

(A-B) Double-Gate H-NTFET device structure and operation (C-D) Double-Gate H-PTFET device structure and operation.
Verilog-A circuit simulation model for TFETs. The energy-delay performance curve of a HTFET 40-stage ring-oscillator, when compared to that of a CMOS ring-oscillator in Fig. 2.9A, shows a cross-over in the energy-delay characteristics. The CMOS ring-oscillator has a better energy-delay compared to the HTFET ring-oscillator at $V_{CC} > 0.65\text{V}$ and the HTFET ring-oscillator has a better energy-delay trade-off at $V_{CC} < 0.55\text{V}$. Other logic gates, such as Or, Not and Xor (which are not shown here) also show a similar cross-over. This trend is consistent with Fig. 2.6 where it has been illustrated that CMOS devices provide better operation at high $V_{CC}$ and HTFETs provide preferred operation at low $V_{CC}$. Fig. 2.9B shows the energy-delay performance of a 32-bit prefix-tree based Han-Carlson Adder has a similar crossover behavior for CMOS and HTFETs.

![Fig. 2.9. Energy-Frequency performance comparison of CMOS and HTFET complementary digital logic - (A) Ring-Oscillator and (B) 32-bit Han-Carlson Adder.](image)

2.3.3 N-Channel Tunnel FET Pass-Transistor Logic

Due to the asymmetric source-drain architecture (Fig. 2.10A-B), HTFETs cannot function as bi-directional pass transistors. Though this may seem to limit the utility of TFETs as access-transistors in SRAM cell design, several SRAM designs have been proposed to circumvent this problem [40, 41]. It is also important to consider a solution for pass-transistor logic, because of the ubiquitous usage of pass-transistors in logic design. We propose using a pass-transistor logic-stack made of n-channel HTFETs, with a n-channel HTFET for pre-charging the output
Fig. 2.10. (A) Asymmetric source-drain architecture for a heterojunction NTFET and (B) Asymmetric $I_D-V_D$ characteristics resulting from source-drain asymmetry.

(Fig. 2.10C). All the HTFETs in the pass-transistor logic-stack can be oriented away from the output, which would allow them to drive their on-current and discharge the output node when the input signals are enabled. During the pre-charge phase, the pre-charge HTFET charges the output to $V_{CC}$, and during the evaluate phase, the HTFET logic-stack evaluates the output based on the inputs.

### 2.3.4 Tunnel FET SRAM Cache

In order to model TFET-based processor architectures, it is important to consider the characteristics of the L1 cache, which is an integral on-chip component of a processor. We use the analytical method implemented in the cache analysis tool CACTI [42], in order to evaluate the energy-delay performance of a TFET-based cache. As discussed in section 2.3.3, in order to overcome the problem of asymmetric conduction in TFETs, a precharge-based pass-transistor mux (Fig. 2.10C) is implemented in CACTI. For the architectural study performed in this chapter, we assume an L1 cache implemented using 6-T SRAM bit-cells with virtual-ground, which was proposed in [41]. A more in-depth discussion on the design and analysis of TFET-based SRAM cells is provided in chapter 3.

To compute the gate delay, CACTI uses the Horowitz approximation [43] given
by:
\[ \tau_{Delay} = \tau_f \sqrt{(\log(V_s))^2 + 2.(\tau_{in}/\tau_f).b.(1-V_s)} \]

where, \( \tau_f = R_{Eff} \times [C_{Load} + C_{Eff}] \) is the time constant, and \( \tau_{in} \) is the input ramp time (b and \( V_s \) are set to 0.5). Effective Switching Resistance (\( R_{Eff} \)) and Effective Output Capacitance (\( C_{Eff} \)) are estimated using transistor switching delay simulations for different load capacitances (\( C_{Load} \)), as described in [44]. This takes into account the effect of TFET Enhanced Miller Capacitance [45] on TFET switching delay. In order to validate the Horowitz model for TFETs, we compare the TFET-inverter delay from the Horowitz analytical expression with the delay estimated using the Verilog-A table-lookup model for different input ramp times (\( \tau_{in} \)), and obtain a good match as shown in Fig. 2.11.

We modified CACTI [42] to implement the 6T TFET SRAM cell design proposed in [41] and evaluated the energy-delay performance of a 32KB L1 cache with a 32-byte block-size and an associativity of 2. Fig. 2.12 shows that a cross-over point similar to that in logic exists for CMOS and TFET-based SRAM L1 caches. Due to the higher \( I_{On}/I_{Off} \) ratio of TFETs, the TFET L1 cache has lower leakage power than the Low-\( V_T \) CMOS L1 cache.
2.4 Architectural Analysis of CMOS and HT-FET cores

2.4.1 Processor Abstraction

The processor frequency is set by a critical path in the logic, which we model using a ring-oscillator chain of NAND gates each driving a FO4 load, as shown in Fig. 2.13A. We observe that this ring-oscillator logic chain is able to model the Vcc-frequency behavior of the Intel Atom processor [46] accurately, as shown in Fig. 2.13B. Further, it was shown in [47] that the power consumption of several representative embedded processors is largely independent of the instruction-stream, and to a first order, dependent only on the VCC and the operating frequency. We use this theory to model the power-frequency of the Atom processor. For a given Vcc, we estimate the switching-energy per transistor using the ring-oscillator. Using the Atom processor’s transistor count of 47 million, and transistor switching activity factor of 30%, we are able to model the power-frequency of the Atom processor accurately, as shown in Fig. 2.13C.

The first order model for estimating processor power holds true provided that a number of the overheads incurred during instruction execution are common across various instructions [47]. This is a reasonable assumption for simpler RISC pipelined processors [48, 46, 49, 50]. In case of processors with more complex in-
struction sets, it may be necessary to use a second-order model in which the power consumption is estimated differently for different instruction classes, as suggested in [47]. A logic standard cell-library is currently being developed to estimate the power consumption in processors with complex instruction sets, also while taking into account the parasitic RC delay resulting from the interconnect.

Based on this ring-oscillator approach, Fig. 2.14 shows the processor-level voltage-frequency coordinates that can be achieved for a HTFET and a CMOS-based processor respectively (minimum frequency is 500 MHz and frequency increases in steps of 125 MHz). These voltage-frequency values are used in the benchmark analysis in section 2.4.2. It is clear from this figure that HTFETs are the preferred device when operating below 1250 MHz. We consider a heterogeneous technology-asymmetric multi-core processor, with TFET processors operating in the 1250-500 MHz frequency range, and CMOS processors operating in the 0.7V to 0.5V range (frequency 1375-500 MHz).
2.4.2 Benchmark Execution

The detailed processor and cache parameters for simulating single-core processors using Simics [51] are shown in Table 2.2. The delay and power numbers for each voltage-frequency pair obtained using circuit simulations are incorporated into the simulation framework. For power analysis, we use a utilization based approach. The utilization is monitored by tracking the execution and stall cycles of the processor using Simics. For the execution cycles, the dynamic energy is computed based on the ring-oscillator approach described in section 2.4.1, whereas leakage power is consumed during both execution and stall cycles (no power-gating is assumed). The cache power models are based on CACTI [42] that incorporates the modifications mentioned in section 2.3.4. We evaluate both single-threaded (SPEC CPU2006 [52, 53]) and multi-threaded (SPLASH [54]) benchmarks. For clarity, we highlight the results from 8 SPEC 2006 and 4 SPLASH benchmarks that capture the major trends observed across the suite.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor Pipeline</td>
<td>Sun’s SPARC based core</td>
</tr>
<tr>
<td>Issue width</td>
<td>1</td>
</tr>
<tr>
<td>Fetch Queue</td>
<td>32</td>
</tr>
<tr>
<td>L1 cache</td>
<td>32KB, 2-way 32B block</td>
</tr>
<tr>
<td>L2 cache</td>
<td>2MB, 8-way 64B block</td>
</tr>
<tr>
<td>Mem. Lat / Baseline Freq.</td>
<td>70 cycles / 2 GHz</td>
</tr>
<tr>
<td>Technology / Voltages</td>
<td>22 nm / $V_{CC} = 0.7V - 0.3V$</td>
</tr>
<tr>
<td>DVFS Interval Period</td>
<td>200,000 Instructions</td>
</tr>
</tbody>
</table>

We then execute various benchmarks (SPLASH benchmarks are executed using a single thread) using (1) an Energy-Aware DVFS policy which seeks to minimize
Fig. 2.15. (A) Processor frequency distribution (B) Normalized Delay and (C) Normalized EDP for Energy-Aware DVFS on benchmarks.

The $E^2$ [55], and (2) a purely IPC-aware DVFS algorithm [56]. The energy-aware DVFS policy monitors if the $E^2$ in a DVFS interval (using the energy and delay incurred in executing 200,000 instructions during the DVFS cycle) is better than the previous interval, and if so, it continues the voltage-frequency (VF) change (either continuing to increase or decrease) - otherwise, the direction of the VF change is reversed. We find that, when using the energy-aware DVFS policy on TFETs, most of the applications spend a significant amount of time (close to 60%) in 1000 MHz to 750MHz range, whereas when using CMOS most of the applications execute in 1375 MHz to 1250 MHz range (Fig. 2.15A). As Fig. 2.14 shows, the relationship between E and $D^2$ for TFET processors is non-linear, and the energy-aware DVFS algorithm sees a significant energy benefit when operating in these frequency ranges (1000 - 750 MHz) with TFETs. Consequently, there is a significant energy-delay benefit (average 50%) when using TFETs over the
baseline CMOS based design (Fig. 2.15C), but with a 40% cost in performance (Fig. 2.15B).

The IPC-aware DVFS algorithm, on the other hand monitors the change in the IPC of the processor and ramps the frequency up or down by 125 MHz when it detects a 5% change in IPC. Fig. 2.16A shows that the degradation in performance is less 12% than compared to baseline CMOS when using IPC-aware DVFS on TFET. Fig. 2.16B shows that the energy reduction is significant when using DVFS on TFETs due to the lower energy of lower frequency modes in TFETs (Energy reduction 26% and ED reduction 18% over baseline CMOS). Further, Fig. 2.16D shows that there is significant $ED^2$ reduction over baseline CMOS (upto 9% $ED^2$ benefit over baseline CMOS) for applications such as bzip, mcf and ocean. These applications have significant L2 miss-rates (shown in Fig. 2.17)
and consequently, the processors spend a lot of time stalling. Thus, by using energy oriented DVFS scaling on TFETs during these stall cycles gives us significant energy advantage when compared to a CMOS based design. We conclude that in heterogeneous-technology asymmetric-performance multi-core processor, single-threaded applications with higher miss-rates are more suited for execution on TFETs with IPC-aware DVFS, since it results in significant $ED^2$ advantage. Pure energy conservation is best achieved by executing the applications on the TFET processor with energy-aware DVFS. Applications such as gcc, perl, sjeng and radix with low cache miss-rates are best executed on the CMOS processor with higher performance.

![Fig. 2.17. L2-cache miss rates for various benchmarks.](image)

Multi-core processors can be used to minimize energy consumption by scaling down the operating frequency and increasing thread-level parallelism in order to regain *iso-performance* to baseline CMOS (1-Core @ 1375 MHz) as shown in Fig. 2.18A. Fig. 2.18B shows the energy consumption compared to baseline CMOS for parallel program execution on 2-Core CMOS and 2-Core TFET, for *iso-performance* to baseline CMOS. When moving from 1 to 2 cores, we observe almost linear performance scaling with the number of cores, that drops the required operating frequency for *iso-performance* below the CMOS-TFET cross-over point. Due to the energy advantage of TFET processors at lower frequencies, TFET processors have a distinct energy advantage in *iso-performance* multi-core execution, giving an energy savings of 70% against single-core CMOS and energy savings of 55% against 2-core CMOS.

The proposed CMOS-TFET hybrid architecture provides additional energy effi-
ciencies for multi-threaded applications by scheduling performance critical threads [57] on high performance CMOS cores and non-critical threads on energy efficient TFET cores. They can exploit imbalance across threads due to application behavior [58].

![Normalized Delay vs Normalized Energy](image)

**Fig. 2.18.** (A) Illustration of normalized energy-delay for iso-performance for OCEAN benchmark and (B) Normalized multi-core execution energy for iso-performance to CMOS @ 1375 MHz.

### 2.5 Conclusions

In this chapter, detailed device simulations have been performed to understand the low-$V_{CC}$ energy-delay benefits of heterojunction TFET-based logic. An abstraction has been proposed to capture the processor-level energy-delay behavior starting from transistor-level Tunnel FET models. We have shown the effectiveness of the hybrid TFET-CMOS core in exploiting inter-application characteristics in multi-programmed workloads. We have also shown that TFET cores become preferable in multi-threaded applications, when conserving energy through increased parallelism. From an Operating System’s (O.S.) perspective, the following scheduling policy can be adopted to take advantage of the Energy-Delay benefits of a heterogeneous TFET-CMOS chip multi-processor:

- Single-threaded application, low-power consumption execution profile: Execute on TFET processor with energy-aware DVFS policy
• Single-threaded application with low cache miss rate, performance-oriented execution profile: Execute on CMOS processor with performance-aware DVFS policy

• Single-threaded application with high cache miss rate, performance-oriented execution profile: Execute on TFET processor with performance-aware DVFS policy

• Multi-threaded application, energy-efficient execution profile with iso-performance to single-core CMOS: Execute on multiple TFET processors at low-$V_{CC}$

• Multi-threaded application, performance-oriented execution profile: Execute on multiple CMOS processors at high-$V_{CC}$

This work has been accepted and presented at the Design Automation Conference (DAC), June 2011, San Diego, CA [59].

The proposed hybrid architecture can also be used to exploit intra-application thread-level characteristics. Using O.S. support, phases can be detected in threads which would be benefited by being scheduled on a CMOS core, and phases that would benefit by being scheduled on a TFET core. This was done in a later work that was accepted and presented at the International Symposium on Low Power Electronics and Design (ISLPED), Aug 2011, Fukuoka, Japan [60].
Variation-Tolerant Ultra Low-Power Heterojunction Tunnel FET SRAM Design

3.1 Introduction

Numerous design techniques have been proposed both at the circuit-level and the architectural-level [61, 62] to enable low-V$_{\text{CC}}$ operation using CMOS digital circuits. Further requirement for energy reduction drives the operation of CMOS digital circuits into sub-threshold operation, thus increasing the sensitivity of the circuit parameters to device-level variation. It also causes exponential increases in delay, causing the circuit operation to be leakage-energy dominated. SRAM bit-cells employing minimum-sized transistors can be particularly vulnerable to device-level variation occurring due to the process flow, intradie as well as interdie [63]. Due to the added sensitivity of minimum sized transistors to variation at the device-level, SRAM bit-cells are most prone to access failures in reduced V$_{\text{CC}}$ operation. Thus, there is a need for robust variation-tolerant SRAM design, capable of sub-300mV operation. Numerous designs have been proposed to address the challenge of sub-threshold operation of CMOS SRAMs [64, 65, 66]. Ultra-low V$_{\text{CC}}$ operation of CMOS SRAM at 160 mV has also been shown [67].
This chapter presents a study of TFET SRAMs from a combined technology-architecture perspective. This study is organized as follows:

1. In section 3.2, we take a closer look at unique features that occur in the electrical characteristics of Tunnel FETs, such as unidirectional conduction and delayed saturation, and analyze their impact on the characteristics of 6T TFET SRAM bit-cells. As an alternative to 6T TFET SRAM bit-cells, we propose a wider design space of 8T and 10T TFET SRAM bit-cells and analyze their figures-of-merit.

2. In section 3.3, we propose a model for studying variation in Ultra-Thin-Body (UTB) TFETs and describe a small-signal variation model suitable for circuit simulations. We extract the $V_{CC}$-min for different CMOS and TFET SRAM bit-cells, and identify a TFET SRAM bit-cell capable of ultra-low $V_{CC}$ operation.

3. In section 3.4, we summarize key conclusions from the SRAM bit-cell benchmarking and variability study performed in this chapter.

### 3.2 SRAM Cell Design and Characterization

#### 3.2.1 TFET and CMOS Device Models

For TFET SRAMs, we use the UTB heterojunction TFET model described in section 2.2.2. We use a CMOS double-gated structure for the study of SRAM bit-cells in this chapter because, as described in section 1.1, multi-gate transistor structures are important for future logic technology generations [68, 8]. The on-current of the simulated DG-MOSFET is calibrated with an experimentally demonstrated Si FinFET, as shown in Fig. 3.1 [68].

Further, we scale the gate length, channel thickness and the oxide thickness of the calibrated Si DG-MOSFET and simulate a highly scaled UTB Si DG-MOSFET with a nearly ideal 60-mV/decade sub-threshold slope. The simulation parameters for the scaled Si DG-MOSFET are shown in Table 3.1. The structure of the simulated DG-MOSFET is identical to the structure of the UTB heterojunction TFET, as shown in Fig. 3.2. Identical structures allow us to capture the impact
of equivalent physical parameter fluctuation on the electrical characteristics of the MOSFET and the TFET, as discussed in section 3.3.

**Table 3.1.** Scaled Si DG-MOSFET simulation parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gate Length ($L_G$) [nm]</td>
<td>40</td>
</tr>
<tr>
<td>Channel Thickness ($T_{Ch}$) [nm]</td>
<td>5</td>
</tr>
<tr>
<td>Oxide Thickness ($T_{Ox}$) [nm]</td>
<td>$2.5 \text{ (Hi}<em>x\text{, }\epsilon</em>{22})$</td>
</tr>
<tr>
<td>Mobility ($\mu$) [cm$^2$/V-sec]</td>
<td>330</td>
</tr>
<tr>
<td>$V_{sat}$ [cm/sec]</td>
<td>$3.6 \times 10^7$ [34]</td>
</tr>
<tr>
<td>SS [mV/decade]</td>
<td>60.6</td>
</tr>
<tr>
<td>$V_T$ [V]</td>
<td>0.45</td>
</tr>
</tbody>
</table>

For the SRAM bit-cell benchmarking study in section 3.2.3 in this chapter, $L_G$ 40nm Si CMOS and heterojunction TFETs have been used. The on-current of the simulated p-channel DG-MOSFET is $1/2$ of the drive current of the n-channel DG-MOSFET discussed here. For simplicity, the p-channel HTFET drive is assumed to be $1/2$ of the n-channel counterpart (p-channel TFETs are discussed in detail in chapter 4).
3.2.2 Impact of TFET Late-Saturation Characteristics on SRAM

It has been shown in section 2.2.2, by comparing the $I_{on}$ vs $I_{on}/I_{off}$ characteristics of a HTFET and a silicon MOSFET, that the heterojunction TFET is a superior device compared to a silicon MOSFET in its sub-threshold region, showing both higher $I_{on}$ as well as higher $I_{on}/I_{off}$ ratio. However, it is also important to consider the $I_{DS}$-$V_{DS}$ characteristics of the HTFET, since the saturation voltage ($V_{DS-Sat}$) plays an important role in determining the noise-margin characteristics of digital circuits. The $I_{DS}$-$V_{DS}$ saturation characteristics from Fig. 2.10B are reproduced and compared with the $I_{DS}$-$V_{DS}$ of the scaled Si DG-MOSFET in Fig. 3.3. The heterojunction TFET behaves like a device with a very low $V_T$ (close to 0V), and hence shows delayed output saturation characteristics. Apart from delayed saturation, the heterojunction TFET also shows uni-directional conduction due to the asymmetric $p-i-n$ structure.

Fig. 3.3. Comparison of $I_{DS}$-$V_{DS}$ characteristics for UTB HTFET and UTB Si DG-MOSFET

Fig. 3.4. Voltage Transfer Characteristic (VTC) comparison for HTFET and Si CMOS
We use the VerilogA Look-up-Table based model described in section 2.3.1 to carry out the SRAM simulations in the rest of this chapter. Because of delayed onset of saturation in the HTFET, the Voltage Transfer Characteristics (VTC) of a HTFET inverter are considerably degraded compared to that of a CMOS inverter, as shown in Fig. 3.4. Further, we consider a 6T TFET SRAM (Fig. 3.5A) with inward-facing access transistors, \textit{(inward} being defined as being able to conduct current from the bit-line into the storage-node of the cell, i.e. the direction in which the arrow points), and a 6T CMOS SRAM (Fig. 3.5B) with similar sized transistors and compare their read-SNMs. We find that read-SNM for the 6T TFET SRAM is considerably degraded compared to the 6T CMOS SRAM (Fig. 3.5C-D). Further, due to uni-directional conduction, the Write-SNM for the TFET SRAM cell shown
in Fig. 3.5A is zero [40].

3.2.3 Design of 8T and 10T TFET SRAM

3.2.3.1 Related Work

As discussed in section 3.2.2, the 6T TFET SRAM with inward-facing access transistors cannot perform a write successfully. It has been shown that a 6T TFET SRAM with either inward or outward-facing access transistors, cannot simultaneously do both read and write [40]. In order to circumvent this limitation, a 6T TFET SRAM with one inward and one outward-facing access transistor has been proposed [41]. However, a virtual-ground write-assist is required to perform a write successfully in this design. Another proposed approach [69] is to use a 6T TFET cell with a cell-ratio ($\beta$) of 0.6 to provide a robust write. This cell has a read-SNM close to zero (because of the low $\beta$ value), and is fundamentally unstable during read. Instead, it relies on the application of a short read-pulse width, relying on the read-dynamic noise margin (DNM) characteristic of the 6T TFET cell, along with a ground-lowering read-assist, to avoid an upset during read operation. Thus, only 6T TFET cells which require a read/write-assist, or cells which are fundamentally unstable during read/write-access have been studied. In this work, we consider the design and characterization of TFET SRAM cells with higher (8T and 10T) transistor counts, and compare them with 6T and 10T CMOS SRAM cells. We do not consider cells which are fundamentally unstable during read/write-access, or cells which require ground-assist to perform a read or write operation ([69, 41]).

3.2.3.2 Read Operation in TFET SRAM cells

Fig. 3.6 shows the read operation for various 8T and 10T TFET SRAM cell configurations. The TFET 8T Transmission-Gate SRAM cell in Fig. 3.6A has both inward and outward-facing TFET access transistors to overcome the problem of unidirectional conduction. Read is performed by enabling the inward-facing TFET access transistors AXL$_{rd}$(AXR$_{rd}$) using the read word-line (WL). The 8T(10T) dual-port SRAM cell ([70, 71, 65]), shown in Fig. 3.6B, has separate read and write access ports. Thus, the read-SNM of the dual-port SRAM is same as its hold-SNM. CMOS Schmitt-Trigger (ST)-based SRAM cells have been proposed in
Fig. 3.6. Read-operation in various 8T and 10T TFET SRAM cell configurations.

[67] and [72]. The TFET ST-based SRAM cells (ST-1 and ST-2) differ from the CMOS counterparts only in the orientation of the access transistors. The TFET ST-1 SRAM cell (Fig. 3.6C) has inward-facing access transistors (AXL<sub>rd</sub>/AXR<sub>rd</sub>), as well as inward-facing feedback transistors (NFL/NFR). Read operation is per-
formed by enabling the inward-facing TFET access transistors AXL\textsubscript{rd}(AXR\textsubscript{rd}) using the word-line (WL), while feedback is provided by the inward-facing NFR(NFL) transistors. The TFET ST-2 SRAM cell (Fig. 3.6D) has outward-facing access transistors (AXL\textsubscript{wr}/AXR\textsubscript{wr}) and inward-facing feedback transistors (NFL/NFR). Read operation is performed by enabling the inward-facing TFET feedback transistor NFL(NFR) using the read word-line (WL), while feedback is provided by the other inward-facing NFR(NFL) transistor.

### 3.2.3.3 Write Operation in TFET SRAM bit-cell

Fig. 3.7 shows the write operation for various TFET SRAM cell configurations. For the TFET 8T Transmission-Gate SRAM cell (Fig. 3.7A), both word-lines (WL/WWL) are enabled during a write. The outward-facing access transistor AXR\textsubscript{wr}(AXL\textsubscript{wr}) drives the cell-node voltage $V_R(V_L)$ to 0, and the inward-facing access transistor AXL\textsubscript{rd}(AXR\textsubscript{rd}) assists in the write by raising the voltage of the complementary cell-node $V_L(V_R)$. For 8T(10T) dual-port SRAM (Fig. 3.7B), the write operation is uni-axial due to the uni-directional conduction property of the outward-facing access transistors. During a write, the cell-node voltage $V_R(V_L)$ is driven to 0 only by the AXR\textsubscript{wr}(AXL\textsubscript{wr}) transistor, while the other transistor AXL\textsubscript{wr}(AXR\textsubscript{wr}) does not assist. The TFET ST-1 SRAM (Fig. 3.7C) also suffers from a uni-axial write operation because all the access transistors face inwards in this cell. In the TFET ST-2 SRAM, both word-lines (WL/WWL) are enabled during a write. The outward-facing access transistor AXR\textsubscript{wr}(AXL\textsubscript{wr}) drives the cell-node voltage $V_R(V_L)$ to 0, and the inward-facing feedback transistor NFL(NFR) assists by raising the voltage of the complementary node.

### 3.2.3.4 Cell Sizing for TFET ST-1 SRAM cell

Cell-sizing has to be studied carefully for TFET SRAM cells in which the write operation is uni-axial (i.e. only one transistor participates in operation). For the 8T and 10T Dual Port SRAM cells, the write-access transistors face outwards and need to have sufficient width for a write operation to be completed unassisted. The cell-sizing in Table 3.2 suffices for the dual-port SRAMs. For the TFET ST-1 SRAM cell, inward-facing access transistors are used for both read, as well as write
Fig. 3.7. Write-operation in various 8T and 10T TFET SRAM cell configurations.

operation. Fig. 3.8A shows the dependence of SNMs on the Pull-up Ratio (PR) @ $V_{CC}$ 300mV, assuming a fixed cell-ratio ($\beta$) of 1. The hold-SNM as well as the read-SNM are sufficiently large even for a very low Pull-up Ratio (PR 0.1) because of the Schmitt-Feedback action of the NFL and NFR transistors (Fig. 3.6C). In
Fig. 3.8. (A) Static Noise Margins vs. Pull-up Ratio (PR) for TFET ST-1 Cell. Note†: Improvement in read-SNM is due to Schmitt-Feedback. Note‡: The weak p-channel is overpowered by the Fwd. biased current of the n-channel access transistor resulting in a strong write. (B) Static Noise Margins as a function of $V_{CC}$ for a Pull-up Ratio of 0.15

fact, Fig. 3.8A also shows that, without feedback and for low PR values, the read-SNM is very low making the cell unstable during reads (this is consistent with the observation made previously [40]).

Write operation can be performed in the ST-1 cell by setting one of the bit-lines to 0, and by enabling the word-line (WL). Since, the feedback transistor is powered by the bit-line supply voltage, the feedback is disabled when the bit-line is set to 0 (Fig. 3.7C). When the pull-up is sufficiently weak (PR < 0.2), write can be performed successfully with a good write-SNM. When the pull-up is strong (PR ≥ 1), the cell retains its data even when the feedback is disabled, causing the uni-axial write to fail. Further, disabling the feedback during write has the unwanted side-effect of disabling Schmitt-Feedback for all the cells which are column-neighbors of the row being written. Fig. 3.8A also shows the hold-SNM, with and without feedback, showing that there is sufficient hold-SNM (> 95mV @ $V_{CC}$ 300mV) even when the feedback is disabled, suggesting that temporary disabling of feedback in the hold-state is not a serious hindrance. Fig. 3.8B shows the read, write and hold-SNM (with and without feedback), as a function of $V_{CC}$ for a Pull-up Ratio of 0.1, illustrating that this sizing strategy is valid at all $V_{CC}$. Thus, using a sizing study for the TFET ST-1 cell, we show that it is possible to take advantage of the Schmitt-Feedback principle to circumvent the problem
of weak read-SNMs in TFET SRAMs, and design an ST-1 SRAM cell capable of unassisted read and write operation. The sensitivity of the read-SNM of the TFET ST-1 cell to device-level variation is explored in section 3.3.2. Section 3.2.4 illustrates a benchmarking study using various SRAM cells discussed here.

3.2.4 Characterization of 8T and 10T TFET SRAM

Table 3.2. Summary of SRAM cell sizing for iso-area comparison (W is the nominal width of a transistor).

<table>
<thead>
<tr>
<th></th>
<th>NL1/</th>
<th>NL2/</th>
<th>PL/</th>
<th>AXL_{Wr}/</th>
<th>AXL_{Rd}/</th>
<th>NFL/</th>
<th>N1/</th>
<th>N3/</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NR1</td>
<td>NR2</td>
<td>PR</td>
<td>AXXR_{Wr}</td>
<td>AXXR_{Rd}</td>
<td>NFR</td>
<td>N2</td>
<td>N4</td>
</tr>
<tr>
<td>6T (CMOS) - 4X Sized</td>
<td>8W</td>
<td>-</td>
<td>4W</td>
<td>4W</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>8T Transmission-Gate</td>
<td>2W</td>
<td>-</td>
<td>W</td>
<td>3W</td>
<td>W</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>8T Dual Port</td>
<td>2W</td>
<td>-</td>
<td>W</td>
<td>3W</td>
<td>-</td>
<td>-</td>
<td>W</td>
<td>W</td>
</tr>
<tr>
<td>10T Dual Port</td>
<td>2W</td>
<td>-</td>
<td>W</td>
<td>2W</td>
<td>-</td>
<td>-</td>
<td>W</td>
<td>W</td>
</tr>
<tr>
<td>Schmitt-Trigger (ST-1)</td>
<td>2W</td>
<td>2W</td>
<td>0.1W</td>
<td>W</td>
<td>2W</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Schmitt-Trigger (ST-2)</td>
<td>2W</td>
<td>2W</td>
<td>W</td>
<td>2W</td>
<td>-</td>
<td>2W</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

3.2.4.1 Iso-Area Condition

In order to compare the SRAM figures-of-merit, the transistor sizing has to be such that an iso-area condition is met. The iso-area requirement means that the memory sub-arrays realized using the candidate SRAM cells, while accounting for single-ended or differential read peripheral circuitry, should have the same area footprint. A sizing strategy has been proposed in [73] in order to study the figures-of-merit for various CMOS SRAM cells. In this paper, we assume that the relative cell sizes of SRAM cells realized using CMOS and HTFETs are comparable. Hence, we adopt a sizing strategy similar to [73], in order to compare the figures-of-merit of various CMOS and TFET SRAM cells. The cell-sizing used for various SRAM cells is shown in Table 3.2. Only the TFET ST-1 SRAM cell has a greatly downsized pull-up transistor because an extremely weak pull-up is necessary for unassisted write operation. This changes the iso-area condition for the TFET ST-1 SRAM by a negligible amount (< 5%) compared to the TFET ST-2 SRAM.

3.2.4.2 Hold and Read Noise Margin Comparison

Fig. 3.9 and Fig. 3.10 compare the hold and read-SNMs of various TFET and CMOS SRAM cells. While CMOS SRAMs exhibit a better SNM at higher V_{CC},
the TFET SRAM cells provide better read-SNM characteristics at the desired low $V_{CC}$ regime, due to their better drive currents at low $V_{CC}$. Among the TFET SRAM cells (Fig. 3.9), ST-1 and ST-2 cells have marginally better hold-SNM than the 8T Transmission-Gate SRAM cell because of the Schmitt-feedback. The ST-1 cell is capable of giving a better hold-SNM than the ST-2 cell, but it is only marginally better in this comparison because the pull-up transistor has been downsized to enable write operations. Fig. 3.10 shows that the read-SNM of the 8T Transmission-Gate SRAM cell is considerably degraded because of the delayed saturation in TFETs, as explained in section 3.2.2. The Schmitt-Feedback in the

Fig. 3.9. Comparison of Hold-Static Noise Margin of various SRAM cells

Fig. 3.10. Comparison of Read-Static Noise Margin of various SRAM cells
ST-2 cell improves the read-SNM by 4x. The read-SNM in the ST-2 cell is better than that of the ST-1 cell because the read-access occurs at a secondary node, $V_{NL}$ (Fig. 3.6D). The ST-1 cell has a downsized pull-up transistor, and the read-access occurs directly at the cell storage-node $V_L$, causing the read-SNM to become lower than that of the ST-2 cell.

3.2.4.3 Write Noise Margin Comparison

Fig. 3.11 shows the WNM characteristics of all the cells being considered. All the TFET SRAM cells, including those with uniaxial write (i.e. driven by only one access transistor), have a write-SNM of atleast 35mV, showing that TFET cells with higher transistor counts can perform unassisted writes unlike the 6T TFET SRAM cells. The TFET ST-2 cell and the 8T Transmission-Gate SRAM cell have the best write-SNM because write is performed using two access-transistors, one facing inwards and one facing outwards. We also observe that the 8T(10T) dual-port SRAM cells have the weakest write-SNM due to the uniaxial write operation. The ST-1 cell has a greatly improved write-SNM due to the use of a very weak pull-up, which can be afforded because its read and hold-SNM are protected by the Schmitt-Feedback. Attempting to improve the write-SNM of the 8T(10T) dual-port SRAM cell by downsizing the pull-up would degrade its hold-SNM further, which is already the weakest among all TFET SRAM cells (Fig. 3.9). This shows
3.2.4.4 Energy-Delay Comparison

We use a 256×256 SRAM array with 50fF bit-line capacitance (estimated using the cache estimation tool-CACTI [42]), to estimate dynamic energy consumption and read-access delay for different SRAM configurations. The word-line drives the access-transistors of 256 bit-cells in a row, and is enabled using an appropriately
sized driver circuit. The time taken to develop a 50mV differential bit-line voltage is used to estimate the read-access delay. The energy consumed by the driver in turning on the access transistors, together with the leakage energy consumption of the 256×256 bit-cell array is used to estimate the dynamic energy for the read-access. Fig. 3.12 shows that the TFET based cells have a lower delay compared to sub-threshold CMOS ($V_{CC} < 500$ mV), whereas the CMOS based cells outperform the TFET cells at higher $V_{CC}$. This is consistent with the observation made in Fig. 2.6. The TFET ST-2 cell has the least delay out of all the cells at low $V_{CC}$ due to its wider read-access transistors.

Fig. 3.13 shows the read-access energy consumption in various SRAM cells. For sub-threshold CMOS, the dynamic-energy is dominated by the leakage-component due to exponential increase in the access delay. The 8T Transmission Gate TFET cell, the TFET ST-2 cell and the TFET 8T(10T) dual-port cell, all have outward-facing access-transistors (to enable write), which have forward-biased p-i-n junctions in the hold-state. These forward-biased access transistors consume a significant amount of p-i-n leakage energy for $V_{CC} > 300$ mV. Only the TFET ST-1 cell does not have this forward-biased leakage because all its access transistors face inwards and are reverse-biased. As a result, for $V_{CC} > 300$mV, the TFET SRAM cells consume more dynamic-energy than their CMOS counterparts, mainly due to p-i-n leakage-energy domination. However, at $V_{CC} < 300$mV, all the TFET SRAM cells show sufficiently low forward-bias leakage to allow significantly energy-efficient operation compared to CMOS.

The conclusion of this benchmarking exercise is that ST-based TFET SRAM cells are the best choice for low-$V_{CC}$ ($< 300$mV) operation because they consume significantly lower energy as well as deliver improved performance compared to iso-area silicon CMOS, which is a direct consequence of the steep sub-threshold characteristic.
3.3 Variation Study

3.3.1 Device-Level Variation Model

It is important to consider process variation in TFETs because the on-current in tunneling-based devices is very sensitive to the tunneling-barrier width. Any source of variation which can affect the effective tunneling-barrier width of the TFET can cause a significant on-current variation. Fig. 3.14 shows the variation model that is used to study the impact of variations in the structure and doping of the TFET.

The TFET structures studied previously [31, 39] do not include a gate-source overlap, which is unavoidable when fabricating UTB side-gated interband tunneling structures shown in Fig. 2.3. Since the source is P++ doped, a positive $V_G$ applied during the operation of the n-channel HTFET creates a depletion region in the gate-source overlap region, as shown in Fig. 3.15A. For the variability study of TFET SRAMs, we assume a nominal gate-source overlap of 2nm. A change in the position of the gate-edge over the source causes a fluctuation of the tunnel-barrier width under the gate, thereby causing on-current fluctuation (Fig. 3.15B). This can be a major cause of on-current variation in side-gated tunneling structures. In this variation model, we also take the impact of quantum confinement of the UTB channel into consideration. Fig. 3.16 shows how the effective band-gap at the source-channel heterojunction interface changes with $T_{Ch}$ due to the effect of structural quantization. These effective band-gaps were computed using a self-consistent Schrodinger-Poisson solver assuming that the channel is placed...
in a potential well (i.e. the oxide) [74]. Fluctuations in T_{Ch} can cause the effective tunneling-width at the source-channel heterojunction interface to change, due to fluctuations in the effective band-gap, thereby causing significant on-current fluctuation.

In order to simplify our analysis, we assume only small fluctuations for the various sources of variation in Fig. 3.14. This allows us to express the variation in I_{On} as:

\[ \delta I_{On} = \frac{\partial I_{On}}{\partial T_{Ch}} \times \delta T_{Ch} + \frac{\partial I_{On}}{\partial T_{Ox}} \times \delta T_{Ox} + \frac{\partial I_{On}}{\partial \phi_M} \times \delta \phi_M + ... \]  \hspace{1cm} (3.1)

Further, we also assume that the sources of variation (\delta T_{Ch}, \delta T_{Ox}, \delta \phi_M, etc.) are statistically independent, which allows us to calculate the variance of I_{On} as:

\[ \sigma_{I_{On}}^2 = \left( \frac{\partial I_{On}}{\partial T_{Ch}} \right)^2 \times \sigma_{T_{Ch}}^2 + \left( \frac{\partial I_{On}}{\partial T_{Ox}} \right)^2 \times \sigma_{T_{Ox}}^2 + \left( \frac{\partial I_{On}}{\partial \phi_M} \right)^2 \times \delta \phi_M^2 + ... \]  \hspace{1cm} (3.2)
We simulated 2000 Monte-Carlo samples of NMOS and HTFET devices in TCAD Sentaurus [33], assuming independent Gaussian distributions for the various sources of variation in Fig. 3.14. The Gaussian distributions used for the variation sources are listed in Table 3.3. The statistical distribution of $I_{On}$ @ $V_{CC}$ 0.5V, obtained through simulation of Monte-Carlo samples in TCAD, compares well with the shape of the distribution predicted by eq. 3.2, both for silicon NMOS (Fig. 3.17F) and for HTFET (Fig. 3.17L), showing the validity of this approximation technique.

### Table 3.3. Sources of Variation for Ultra-Thin-Body Device.

<table>
<thead>
<tr>
<th>Variation</th>
<th>$N(\mu, \sigma)$</th>
<th>$\mu$</th>
<th>$3\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel Thickness, $T_{Ch}$ [nm]</td>
<td>5</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Oxide Thickness, $T_{Ox}$ [nm]</td>
<td>2.5</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Source Doping (HTFET) cm$^{-3}$</td>
<td>4.5x10$^{19}$</td>
<td>5x10$^{18}$</td>
<td></td>
</tr>
<tr>
<td>Gate Work Func., $\phi_M$ (HTFET) [eV]</td>
<td>4.85</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>Gate Work Func., $\phi_M$ (NMOS) [eV]</td>
<td>4.48</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>Left Gate Edge [nm] (w.r.t channel center)</td>
<td>-20</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Right Gate Edge [nm] (w.r.t channel center)</td>
<td>+20</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Our basic assumption for studying variation is that deviations from the nominal device parameters are always small, as shown in Table 3.3, so that eq. 3.1 is valid. Based on this assumption, we calculate the variation-coefficients for $I_{On}$ – $\partial I_{On}/\partial T_{Ch}$, $\partial I_{On}/\partial T_{Ox}$, etc. – from TCAD simulations. Figs. 3.17A-E show the variation-coefficients that are extracted for an UTB NMOS device using TCAD simulations, and Figs. 3.17G-K show the variation-coefficients that are extracted for an UTB HTFET device. Using these variation coefficients, we are able to extend the Verilog-A table look-up model discussed in section 2.3.1 to study the impact of device-level variation on circuit characteristics. In this study, we have only modeled the on-current fluctuations using eq. 3.1. Since the read-SNM of SRAM bit-cells is mainly dependent on the on-current, this model suffices for analyzing the impact of variation on SRAM circuit characteristics.

Variation in TFETs has been studied previously considering only two parameters - $T_{Ch}$ and $T_{Ox}$ [75]. However, there are also other prominent sources of variation that occur in a side-gated TFET, which are taken into consideration in our model (Fig. 3.14 and Table 3.3). Fig. 3.18 (based on eq. 3.2) shows a breakdown of the contribution of various sources of variation in TFETs, to the total on-current variance ($\sigma I_{On}^2$). It shows that the gate-source overlap (Fig. 3.15) can
be a significant source of variation in an UTB TFET, for $V_{CC} \geq 500$ mV. Our model also shows that $T_{Cn}$ fluctuation (Fig. 3.16), quantum effects included, can also be a major source of variation, for $V_{CC} \leq 300$ mV.

Fig. 3.19 compares the $\% \sigma_{I_{On}}/I_{On}$ change in CMOS and TFETs due to variations. It can be seen that TFETs are in general prone to variations, where-as
Fig. 3.18. Components of on-current variance in silicon NMOS and HTFET due to various sources of variation @ different V\textsubscript{CC}

Fig. 3.19. Comparison of % \( \sigma \frac{I_{on}}{I_{on}} \) for silicon NMOS and HTFET @ different V\textsubscript{CC}

CMOS is prone to variation only in the sub-threshold region. Thus, it is prudent to compare how variations impact the read-write SNMs in HTFET and sub-threshold CMOS SRAMs. The following sub-section summarizes the impact of variation on SRAM read-write noise margin characteristics using the model described in this section.

3.3.2 Monte-Carlo Simulation of Read-Write Noise Margins

Monte-Carlo simulation at the circuit-level for TFET-based SRAMs was shown in [69] assuming only one source of variation (T\textsubscript{Ox}). We perform Monte-Carlo simulation at the SRAM circuit-level using our proposed small-signal variation model assuming all the possible sources of variation. The read-failure probability
Fig. 3.20. Probability of read-upset for various CMOS and TFET SRAM bit-cell configurations is defined as [73]:

\[ P_{\text{read-upset}} = \Pr\{\text{read-SNM} < kT\} \]

where, \( kT = 26\text{mV} \) at 300K. We generate one-thousand Monte-Carlo samples for various CMOS and TFET SRAM cells, and estimate the mean and sigma of the read-SNM at different voltage points. Using these estimates, we plot the the read-upset probability as a function of \( V_{\text{CC}} \) for different SRAM cells, as shown in Fig. 3.20. The \( V_{\text{CC}}\)-min is defined as the voltage for which \( P_{\text{read-upset}} \) is \( \leq 10^{-9} \). The CMOS ST-2 cell has the best read \( V_{\text{CC}}\)-min of 134mV among the CMOS SRAM cells because of its improved variation tolerance [73]. The 8T Transmission-Gate SRAM had degraded read-SNM due to the delayed saturation in TFETs (Fig. 3.10). In addition, TFETs are prone to variation as discussed in section 3.3.1. As a result, the 8T Transmission-Gate SRAM shows a very high probability of upset even when the \( V_{\text{CC}} \) is increased, showing its unsuitability for low-\( V_{\text{CC}} \) applications. The TFET ST-1 SRAM shows improvement in the read-upset probability compared to the 8T TFET SRAM due to the use of the Schmitt-Feedback. However, due to its weak pull-up, the cell is still prone to variation-induced upsets,
causing its $V_{CC}$-min to be large compared to the CMOS SRAM cells. In contrast, the TFET ST-2 cell, shows sufficient variation tolerance and also shows a $V_{CC}$-min of 124mV, showing the suitability of this cell for ultra low-$V_{CC}$ operation.

Fig. 3.21A shows the $V_{CC}$-min for different cell configurations. Fig. 3.21B-C show a comparison of the dynamic energy and leakage power consumption for different SRAM bit-cell configurations operating at their respective $V_{CC}$-min, using the 256×256 SRAM array discussed in section 3.2.4.4. The TFET ST-2 cell at its $V_{CC}$-min provides 240× lower dynamic energy and 10× lower leakage power consumption compared to the CMOS ST-2 cell operating at its $V_{CC}$-min. At the same time, as explained in section 3.2.4, the TFET ST-2 cell has far better performance than the sub-threshold CMOS ST-2 cell, due to the better drive currents of HTFETs in the low $V_{CC}$ regime.
3.4 Conclusion

In this chapter, we have covered a wide design space of SRAM cells, and used additional transistors (8T and 10T) to overcome the challenges of TFET-based SRAM cell design. We show that, among area-equivalent 8T and 10T bit-cells, Schmitt-Trigger-based (10T) ST2 bit-cell beats CMOS bit-cells in energy, performance, as well as static-RNM in sub-300 mV operation, making it a very attractive low-$V_{\text{CC}}$ alternative compared to sub-threshold CMOS. We also proposed a small-signal variation model to analyze the impact of device-level variation on the stability characteristics of SRAM cells, and showed that Schmitt-Trigger-based TFET SRAM cells have sufficient variation tolerance to allow ultra-low $V_{\text{CC}}$-min operation. This work has been accepted and presented at Nanoscale Architectures (NanoArch), June 2011, San Diego, CA [76].
Chapter 4

Low Power Loadless-4T SRAM cell based on Degenerately Doped Source (DDS) In$_{0.53}$Ga$_{0.47}$As Tunnel FETs

4.1 Introduction

In this chapter we discuss the effect of degenerate source doping on the sub-threshold slope (SS) characteristics of n-channel and p-channel In$_{0.53}$Ga$_{0.47}$As Interband tunnel FETs. We then illustrate the design and operation of a Loadless-4T SRAM cell which is based on these dissimilar SS characteristics.

4.2 Source- Doping Dependant Sub- Threshold Slope Characteristics

The P-N junction between the source and the channel serves as the tunnel barrier in the on-state of the TFET. In order to achieve high tunneling on-current, the source needs to be degenerately doped preferably for the p-type TFET as well as the n-type TFET, so that the width of the tunnel barrier is reduced. $I_D-V_G$ simulations, shown in Figure 4.1, for degenerately doped source InGaAs NTFET and PTFET
show a sub-$\frac{kT}{q} \ln(10)$ SS for the NTFET, and a $\frac{kT}{q} \ln(10)$ temperature-dependent SS for the PTFET. In order to explain the dissimilar SS characteristics, it can noted that the electronic-structure of the conduction and the valence bands for In$_{0.53}$Ga$_{0.47}$As is different, and is reflected in the form of different Density-of-States for the conduction and the valence bands, $N_C = 2.2 \times 10^{17}$ cm$^{-3}$ and $N_V = 1.5 \times 10^{19}$ cm$^{-3}$. The conduction and the valence band offsets in the degenerately doped source regions are obtained from simulation to be, $E_F - E_C = 0.6$ eV for DDS In$_{0.53}$Ga$_{0.47}$As PTFET and $E_V - E_F = 0.055$ eV for DDS In$_{0.53}$Ga$_{0.47}$As NTFET, as shown in Figure 4.2.

The carrier populations that participate in inter-band tunneling are the minority carriers, namely holes in the degenerately doped N++ source of the PTFET,
and electrons in the degenerately doped P++ source of the NTFET. The population of these minority carriers is temperature dependent and shows a $\frac{kT}{q} \ln(10)$ temperature dependence. In the case of the NTFET, the temperature-dependent minority electron carrier population is restricted to the energy-region between $E_V$ and $E_F$ which is only 0.055 eV wide due to the small source-region degeneracy (Figure 4.2A) - In other words, the temperature dependent minority electron carrier population is filtered away by the valence band edge. However, in the case of the PTFET, the temperature-dependent minority hole carrier population is restricted to the region between $E_F$ and $E_C$ which is 0.6 eV wide due to a large source-side degeneracy (Figure 4.2B) - A very small portion of the temperature dependent minority hole carrier population is filtered away by the conduction band edge. Consequently, the DDS PTFET SS characteristic, which is a function of the minority carrier population, shows a temperature dependence as shown in Figure 4.1B. Thus, the amount of degeneracy in a heavily doped source-region can be explained to be the cause of the sub-$\frac{kT}{q} \ln(10)$ SS characteristic of In$_{0.53}$Ga$_{0.47}$As NTFET and the $\frac{kT}{q} \ln(10)$ SS characteristic of the In$_{0.53}$Ga$_{0.47}$As PTFET.

4.3 A 4T-Loadless SRAM Cell Based on DDS III-V TFETs

We propose the design of a loadless-4T SRAM cell taking advantage of the dissimilar SS characteristics of the DDS In$_{0.53}$Ga$_{0.47}$As NTFET and the PTFET. In a loadless-4T SRAM cell, the PMOS load transistors are removed and the NMOS access transistors are replaced by PMOS. In the hold state, shown in Fig. 4.3A, the bit-lines and the word-lines are set to $V_{CC}$, thereby causing the PMOS access transistors to be in the off-state. In order for the 4T cell to remain stable, it is necessary for the off-current of the PMOS access transistor (I-Off-A) to be larger than the off-current of the NMOS drive transistor (I-Off-n) (Fig. 4.3B). In the case of CMOS loadless-4T SRAM cells, this is done by adjusting the $V_T$ [77], so that the PMOS has a higher off-state leakage than the NMOS. In the case of DDS In$_{0.53}$Ga$_{0.47}$As TFETs, the ratio of off-currents between the PTFET access transistor and the NTFET drive transistor is $10^5$, as shown in Fig. 4.3B, due to
their asymmetric SS characteristics. The read operation is shown in Fig. 4.3C, where the voltage of the word-line is lowered to 0. The PTFET access transistor is turned on, thereby conducting current inward to discharge the bit-line. During the write operation, shown in Fig. 4.3D, both the bit-line and the word-line are set to zero, thereby decoupling the cell, and allowing a write to occur through the inward conducting PTFET.

We compare the access-delay of a TFET Loadless-4T SRAM cell with that of the access-delay of a CMOS Loadless 4-T SRAM cell (Figure 4.4). At low $V_{CC}$, the access-delay for the TFET Loadless-4T SRAM cell is lower compared to CMOS Loadless-4T SRAM cells, because of the higher on-current provided by the sub-$\frac{kT}{q}$ SS of the NTFET drive transistor for a given $V_{CC}$.

Due to the absence of a load transistor, a Loadless-4T SRAM cell suffers from a write-upset issue as shown in Figure 4.5A. When a write is performed on a column-neighbour of the cell, the bit-line is forced to zero while the word-line
remains at $V_{CC}$, thereby decoupling the cell and causing the cell state to be lost through capacitive discharge. Due to the low leakage of the NTFET, it takes longer for the SRAM cell to discharge, thus increasing the time-to-upset for the TFET loadless-4T SRAM cell (Figure 4.5B).

### 4.4 Conclusion

This work resulted in a better understanding of the energy-filter in TFETs which produces a sub-$kT/q$ steep sub-threshold slope characteristic. The challenge of realizing a III-V p-channel TFET with both good on-current performance and steep SS is explained – to produce good on-current, the source region has to be degenerately doped n-type, which in turn destroys the sub-$kT/q$ sub-threshold slope characteristic. This result holds for III-V homojunction as well as heterojunction
p-channel TFET.

In chapters 2 & 3, a III-V heterojunction p-channel TFET with a lightly doped InAs source, described in Figs. 2.8C & 2.8D, is used for logic and SRAM simulations. The light source doping ($1 \times 10^{18}$ cm$^{-3}$) ensures that the amount of degeneracy in the source is limited, thereby ensuring a sub-$kT/q$ sub-threshold slope. However, the on-current performance of the heterojunction p-channel TFET (with lightly doped source) is reduced compared to the heterojunction n-channel TFET (with heavily doped source), which is taken into consideration in the circuit simulations.

Further, in this chapter, a TFET loadless-4T SRAM cell is proposed taking advantage of the large off-state leakage difference resulting from the dissimilar SS characteristics of DDS PTFET and NTFET. The proposed TFET loadless-4T SRAM does not require bi-directional access-transistors, and can be implemented using inward facing PTFET access-transistors. The TFET Loadless-4T SRAM cell shows lower delay at low $V_{CC}$ due to improved on-current provided by the steep SS of the NTFET drive transistor.

This work has been presented at the Device Research Conference (DRC), July 2010, South Bend, In. [78]. The work on the TFET Loadless-4T SRAM has also resulted in a Provisional Patent Application being filed with the USPTO.
5.1 Introduction

A Single Electron transistor (SET) is a transistor in the limit of scaling with a very small feature size, and hence very small self capacitance. Thus, it can be turned On and Off using very few electrons, i.e. intrinsically very little charge is necessary to operate the device. Given, the trend that leakage energy dominates the energy consumption of scaled transistors; it is useful to look at the energy-delay tradeoff of an SET device in order to understand what kind of challenges face us in order to reach closer to the theoretical minimum energy limit for switching logic circuits [79].

The rest of this chapter is organized as follows:

1. In section 5.2, we derive an analytical expression for the peak-to-valley current ratio of a Single Electron Transistor. We also show how the quantization energy ($E_Q$) of the SET quantum-dot can be included into this derivation.
2. In section 5.3, we construct a design space for Single Electron Transistors, based on the physical dimensions and electrostatic properties of the SET.

3. In section 5.4, we compare two logic styles for SETs - Complementary Gate logic and Binary Decision Diagrams with Sense Amplifier (BDD-SA) logic. We also benchmark SET logic circuits against 22nm and 16nm silicon CMOS logic circuits.

4. In section 5.5, we summarize the results from this chapter and draw conclusions.

5.2 Analytical Modeling of Single Electron Transistors

5.2.1 Device Electrostatics

We assume a simple symmetric three terminal SET device shown in Fig. 1.6A for deriving the electrostatics of the SET. The capacitance coupling between the gate and the quantum-dot is $C_G$, and the capacitances between the source/drain and the quantum-dot are, $C_S$ and $C_D$ ($C_S = C_D$ due to symmetry assumption). The self capacitance $C_\Sigma$ of the quantum-dot is equal to the sum of the various coupling capacitances ($C_\Sigma = C_G + C_D + C_D$). The quantum-dot is isolated from the source and drain contacts by tunnel barriers with resistance $R_T > h/q^2$. For subsequent derivations, the source of the device in Fig. 1.6A is grounded ($V_{Source} = 0V$).

When a voltage $V_{Drain}$ is applied to the drain contact (for $V_{Gate} = 0V$, $V_{Source} = 0V$), the potential of the quantum-dot is increased by a fraction proportional to the coupling capacitance $C_D$ between the drain and the quantum-dot. This change in potential due to $V_{Drain}$ is given by: $C_D/C_\Sigma \times V_{Drain}$. For the applied drain bias of $V_{Drain}$, the potential of the quantum-dot when the first Coulomb oscillation peak is reached is given by: $q/2C_\Sigma + V_{Drain}/2$. Thus, for an applied drain bias of $V_{Drain}$, the amount of gate voltage that needs to be applied in order to reach the first Coulomb oscillation peak can be derived, as shown in equations (5.1) - (5.3).
As shown in Fig. 5.1, the quantum-dot potential changes by $q/C$ between two consecutive current peaks. Due to the symmetry of the device, the valley occurs between the two Coulomb oscillation peaks. Thus, the quantum-dot potential at the valley must be $q/2C$ less than the quantum-dot potential at the first Coulomb peak as shown in Fig. 5.1. In order to reach the first valley, the gate voltage that needs to be applied can be derived as shown in (5.4)- (5.6).

$$
\frac{C_G}{C}V_{Gate}(I_{peak}) = \frac{1}{2} \left( \frac{q}{C} + V_{Drain} \right) - \frac{C_{Drain}}{C}V_{Drain} \tag{5.1}
$$

$$
V_{Gate}(I_{peak}) = \left( \frac{q}{2C} + \frac{V_{Drain}}{2} - \frac{C_D}{C}V_{Drain} \right) \frac{C}{C} \tag{5.2}
$$

$$
V_{Gate}(I_{peak}) = \frac{q}{2C} + \frac{V_{Drain}}{2} \tag{5.3}
$$

$$
\frac{C_G}{C}V_{Gate}(I_{Valley}) = \frac{1}{2} \left( \frac{q}{C} + V_{Drain} \right) - \frac{C_D}{C}V_{Drain} - \frac{1}{2} \cdot \frac{q}{C} \tag{5.4}
$$

$$
V_{Gate}(I_{Valley}) = \left( \frac{V_{Drain}}{2} - \frac{C_D}{C}V_{Drain} \right) \frac{C}{C} \tag{5.5}
$$

$$
V_{Gate}(I_{Valley}) = \frac{V_{Drain}}{2} \tag{5.6}
$$
According to (5.3) and (5.6), for any \(V_{\text{Drain}}\), the amount of gate swing required to start from \(I_{\text{Valley}}\) and reach \(I_{\text{Peak}}\) is given by \(q/2C_G\).

5.2.2 Analytical Expressions for Peak and Valley Currents

\[
\frac{\partial P_i(t)}{\partial t} = \sum_{j \neq i} \left[ \Gamma_{ij} P_j(t) - \Gamma_{ji} P_i(t) \right]
\] (5.7)

The description of electron currents resulting from single electron tunneling is well understood [80, 35] and is best described as a Stochastic Markov Chain process, as shown in Fig. 5.2A, where each state indicates the charge state of the device. The Master Equation (5.7) is used to solve for the occupation probabilities of each state (\(P_i(t)\) is the time-dependent probability that the device has charge-state \(i\)). This is the most general description of single electron tunneling and is used in well known Monte-Carlo simulators for SETs such as SIMON [35]. A commonly used approximation though, is to model the process of tunneling as a birth-death
process (Fig. 5.2B), where the charge state of the device changes between $N$ and $N + 1$ electrons only. Analytical expressions for the $I_D$-$V_G$ curve of an SET device have been derived previously \[81, 82\] based on the birth-death process model.

Since the aim is to obtain simple expressions for peak and valley currents, we start by considering only two states, $N=0$ and $N=1$, shown in Fig. 5.2, and the tunneling rates $\Gamma(0 \rightarrow 1)$ and $\Gamma(1 \rightarrow 0)$ between these two states. By solving the Master Equation (5.7), it is observed that these two tunneling rates cause the first Coulomb oscillation peak for a positive $V_G$ sweep, as shown in Fig. 5.3. Further, the tunneling rates between the states $N = 0$ and $N = -1$ (\(\Gamma(0 \rightarrow -1)\) and \(\Gamma(-1 \rightarrow 0)\)) cause the first Coulomb oscillation peak for a negative $V_G$ sweep. Thus, the tunneling rates contributing to the first $I_{\text{Valley}}$ current are the four tunneling rates: $\Gamma(0 \rightarrow 1)/\Gamma(1 \rightarrow 0)$ and $\Gamma(0 \rightarrow -1)/\Gamma(-1 \rightarrow 0)$. Based on the electrostatics derived in section 5.2.1, the first $I_{\text{Valley}}$ occurs at $V_{\text{Gate}} = V_{\text{Drain}}/2$. Thus, approximate analytical expressions for the contribution of the tunneling rates $\Gamma(0 \rightarrow 1)/\Gamma(1 \rightarrow 0)$ and $\Gamma(0 \rightarrow -1)/\Gamma(-1 \rightarrow 0)$ to $I_{\text{Valley}}$ and $I_{\text{Peak}}$ can be derived, as shown in equations (5.8) - (5.9). The detailed derivations of the simplified expressions for $I_{\text{Valley}}$ and $I_{\text{Peak}}$ are given in Appendix A.
A comparison of the values for $I_{\text{Peak}}$ and $I_{\text{Valley}}$ obtained through Monte-Carlo simulation in SIMON and those computed using the analytical approximation equations (5.9) and (5.8) is shown in Fig. 5.4, and shows the validity of this approximation. According to (5.8), $I_{\text{Valley}}$ reduces exponentially as $E_0$ (the self-charging energy of the quantum-dot) is increased, which happens as the SET quantum-dot’s physical dimension is scaled down. This clarifies the role that $E_0$ plays in the operation of an SET. The self-charging energy ($E_0$) opens up an energy-gap in the energy state distribution of the quantum-dot allowing us to turn the device On and Off. Thus, scaling down the physical dimension of the SET quantum-dot causes the energy separation $E_0$ to increase linearly, causing the $I_{\text{Valley}}$ to decrease exponentially.
5.2.3 Modeling Quantization Effect

In order to operate SETs at room temperature, the quantum-dot’s dimension needs to be scaled down so that \( E_0 \approx 8.0.300 \). However, as the device is scaled down, there is an increase not only in the self-charging energy (\( E_0 \)) but also in the quantization energy (\( E_Q \)). In order to compare \( E_0 \) and \( E_Q \) as the device dimension is scaled, the quantum-dot is approximated as a sphere located in an infinite potential well, as suggested in [83]. Based on these assumptions, the self charging energy (\( E_0 \)) and the quantization energy (\( E_Q \)) are modeled by equations (5.10) and (5.11).

\[
E_0 = \frac{e}{2.\pi.\epsilon_{\text{silicon}}.d} \quad \text{where } \epsilon_{\text{silicon}} = 11.7 \quad (5.10)
\]

\[
E_Q = \frac{1}{2.m^*.(\frac{\hbar.\pi.N}{d})^2} \quad \text{where } m^*_S = 0.26 , N = 1 \quad (5.11)
\]

As shown in Fig. 5.5, \( E_0 \) and \( E_Q \) have been plotted as functions of the quantum-dot dimension [83]. The quantum-dot diameter (d) such that it’s self-energy (\( E_{Q-Dot} = E_0 + E_Q \)) is large enough for Room-Temperature operation (\( E_{Q-Dot} > 8.0.300 \)), is found to be 3nm and below. At this dimension, the main contributor to \( E_{Q-Dot} \) is the quantization energy \( E_Q \). Thus, there is a need to take quantization energy (\( E_Q \)) into consideration when modeling room temperature SETs. Recently, a 2nm-4nm Room Temperature-Single Electron Transistor (RT-SET) has been demonstrated [19], showing that indeed the quantization energy plays a significant role in enabling RT operation in larger size quantum dots.

Many models have been proposed to include quantization into the SET tunneling rate equations, and these have been validated against experimental data [84, 85]. However, the aim of this chapter is to model the dependence of \( I_{Peak}/I_{Valley} \) ratio, and the peak-to-valley \( V_{Gate} \) swing, on the physical dimensions and electrostatics of an SET device. By recognizing that the self-charging energy (\( E_0 \)) opens up an energy-gap in the energy state distribution of the SET quantum-dot, and that the quantization energy (\( E_Q \)) adds to this energy gap, equations (5.8) and (5.3) can be modified to obtain equations (5.12) and (5.13) for \( I_{Valley} \) and \( V_{Gate}(I_{Peak}) \) in the quantized case.
\[ I_{valley} \approx \frac{1}{q.R_T} \left( E_0 + E_Q - q.V_{Drain} \right) \left( E_0 + E_Q + q.V_{Drain} \right) \frac{2.(E_0 + E_Q)}{2.(E_0 + E_Q)} \cdot e^{-\frac{E_0}{2kT}} \cdot e^{-\frac{E_Q}{2kT}} \cdot e^{-\frac{q.V_{Drain}}{2kT}} \]

\[ V_{Gate}(I_{Peak}) = \frac{1}{2} \cdot \frac{e}{C_G} \cdot \left( 1 + \frac{C_{\Sigma}}{C_Q} \right) \]

The approximation that is used in the quantized case, is to model a SET with quantized energy levels as a classical SET with an increased energy gap (i.e. effective Coulomb gap = \( E_0 + E_Q \)). We clarify that our approach is only an approximation, because the SET rate equations in Appendix A are for a continuous distribution of energy levels in the quantum-dot, and are not directly applicable for an SET with quantized energy levels (precise tunneling rate equations for discrete energy levels are shown in [84, 85]). In order to check the validity of the approximations, the discrete energy-level simulation model in the Monte-Carlo simulator-
SIMON is used. Discrete energy levels are specified in SIMON at $+E_Q/2$ eV and $-E_Q/2$ eV, where $E_Q$ is the quantization energy for a given dot dimension. SIMON models the energy level broadening as a Gaussian function, and hence provides the height (H) and width (W) of broadened energy levels as tunable parameters. For validation of eq. (5.12), the energy level broadening is considered to be minimal, and hence set the Width (W) parameter to 1meV. Since the broadening is assumed to be minimal, the peak current is limited mainly by the tunnel barrier resistance, and hence can be considered to be the same as in eq. (5.9). The height (H) is set such that the current peak for discrete energy level simulation equals that in eq. (5.9). This is reflected in Fig. 5.6, in which $I_{\text{Peak}}$ is the same for the analytical approximation as well as the Monte-Carlo simulation. Fig. 5.6 shows that $I_{\text{Valley}}$ estimated using eq. (5.12) follows the same trend as $I_{\text{Valley}}$ obtained from SIMON, and is off by less than one order of magnitude.
5.3 SET Design Space Exploration

5.3.1 Description of SET Design Space

Eq. (5.12) \( I_{\text{Valley}} \) of an SET and eq. (5.13) \( V_{\text{Gate}}(I_{\text{Peak}}) \) of an SET enable us to describe a design space for SET devices, with the device physical dimensions and electrostatic properties as the parameters. This sub-section provides a description of this design space. Based on Fig. 5.5, the quantum-dot diameter should be 3nm or below for room temperature operation. Since eq. (5.12) for \( I_{\text{Valley}} \) is valid only when \( E_0 + E_Q - qV_{\text{Drain}} \gg k.T \), the quantum-dot diameter is chosen to be in the range of 1.5nm to 2.5nm for our design space (so that \( E_0 + E_Q \) is sufficiently large for Eq (5.12) to hold). Since the focus is on the application of the SET as an ultra-low power device, \( V_{\text{Drain}} \) voltages up to 200 mV are considered for room temperature operation. Using Eqs. Eq (5.13) and Eq (5.12), the \( I_{\text{Peak}}/I_{\text{Valley}} \) ratio can be plotted as a function of the quantum-dot diameter and the applied drain bias. The surface plot in Fig. 5.7A shows that \( I_{\text{Peak}}/I_{\text{Valley}} \) ratio increases by four orders of magnitude as the quantum-dot diameter is reduced from 2.5nm to 1.5nm.

\[
V_{\text{Gate}}(I_{\text{Peak}} \rightarrow I_{\text{Valley}}) = \frac{1}{2}(E_0 + E_Q) \frac{C_G}{C_\Sigma} \tag{5.14}
\]

The gate voltage-swing for \( I_{\text{peak}} \)-to-\( I_{\text{valley}} \), given by eq. (5.14), is dependent on the quantum-dot diameter (which determines \( E_0 + E_Q \)), and the electrical
parameter $C_G/C_Σ$, which is called the gate-control ratio. The gate-control ratio is an electrostatic parameter of the SET device which indicates how well the gate controls the potential of the quantum-dot. The contour plot in Fig. 5.7B shows $V_{\text{Gate}}(I_{\text{Peak}} \rightarrow I_{\text{Valley}})$ as a function of the quantum-dot diameter and the gate-control ratio. This contour plot shows that the gate voltage-swing increases as the quantum-dot diameter scales down, and worsens when the gate-control ratio is lower.

It is useful from a circuit-design standpoint to combine the plots in Fig. 5.7 and plot the switching-slope (SS) characteristic of an SET, as a function of the quantum-dot diameter and applied drain bias, for different gate-control ratios. Fig. 5.8A shows the SS contours for a SET with a high gate-control ratio ($C_G/C_Σ = 7/10$), as a function of the quantum-dot diameter (1.5nm to 2.5nm) and applied drain bias up to 200 mV. Fig. 5.8B shows the same for a SET with a poor gate-control ratio ($C_G/C_Σ = 4/10$). The observation from Fig. 5.8A, is that even though $I_{\text{Peak}}/I_{\text{Valley}}$ ratio increases as the quantum-dot diameter scales down, $\Delta V_{\text{Gate}}(I_{\text{Valley}} \rightarrow I_{\text{Peak}})$ also increases, thus keeping the switching slope nearly constant. Since we are interested in transistor operation at 200 mV, from Fig. 5.8A, it can be observed that the largest quantum-dot diameter which can operate at $V_{\text{Drain}} = 200$ mV with a SS close to 200 mV/decade is 2nm ($E_0 = 0.123$ eV, $E_Q = 0.361$
eV). SETs with quantum-dot diameter larger than 2nm have a SS larger than 200 mV due to a degraded $I_{\text{Peak}}/I_{\text{Valley}}$ ratio. Such an SET (diameter 2nm, gate-control ratio 7/10) can be turned On and Off with a current ratio of 10, by a $V_{\text{Gate}}$ swing whose magnitude equals the magnitude of the applied drain bias of 200 mV. Thus, it is possible to build logic gates with SET devices having such characteristics.

Fig. 5.8B, it can be observed that when the gate-control ratio is low, the SS for a quantum-dot of any dimension operating at a drain bias of 200 mV is degraded and in excess of 250 mV, mainly due to the poor control of the gate over the quantum-dot potential. The consequence of this is that, it is not possible to build logic gates using using SET devices with poor gate-control, because the $V_{\text{Gate}}$ swing necessary to turn the devices On and Off is greatly in excess of the applied drain bias (200 mV). Thus, in the case of SETs with poor gate-control, external amplification is necessary to increase the gate swing, in order to drive the next logic stage.

5.3.2 SET- Device Design Case Study

The architecture of a programmable SET device with split-gate tunable tunnel barriers was proposed in [86]. Such a device structure has been simulated in TCAD Sentaurus [33] using a modulation-doped SiGe/strained-silicon heterostructure, as shown in Fig. 5.9. Device simulation performed at 4K shows a 2D electron-gas (2DEG) in the strained-silicon layer at the bottom as shown in Fig. 5.9.

Fig. 5.10 shows snapshots of this device in it’s various modes-of-operation. When -200 mV bias is applied to the split-gates, depletion regions are formed in the strained-silicon layer underneath, as shown in Fig. 5.10A. Fig. 5.10A shows a quantum-dot being formed in the 2DEG in the strained-silicon layer, which is isolated from the source/drain contacts by the depletion regions formed under the split-gates. The depletion regions act as tunneling barriers causing a tunneling current to flow when a drain bias is applied, as shown in Fig. 5.10B. In the simulated structure, a tunneling current of 0.4 nA flows for the applied drain bias of 1mV at 4K, thus giving the depletion-region tunnel-barrier resistance to be 1.25 MΩ.

Using AC analysis in TCAD Sentaurus, the control-gate-to-quantum-dot ($C_G$) coupling capacitance was found to be 17.3 aF, the split-gate-to-quantum-dot ($C_{WG}$)
coupling capacitance was found to be 20.7 aF and the drain(source)-to-quantum-dot coupling capacitances ($C_D$ and $C_S$) were found to be 3.7 aF. Based on these parameter estimates, the device is modeled in the Monte-Carlo simulator SIMON (Fig. 5.11A) and Coulomb oscillations shown in Fig. 5.11B are obtained.

Fig. 5.10C shows that when the split-gate bias is -175 mV, a depletion region is not formed under the split-gates causing the structure to behave like a conducting wire (i.e. a short). Fig. 5.10D shows that when the split-gate bias is -225 mV, a very large depletion region is formed under the split-gates causing the structure to behave like a non-conducting wire (i.e. an open). Fig. 5.12 shows the SEM image of an experimentally realized split-gate SET [21]. The electrical characteristics of this device shown in Fig. 5.13, demonstrate the operation of this device in three modes - Short, Open and Coulomb blockade. Thus, using rigorous TCAD simulations, the functionality of a programmable SET device, which can behave like an open, a short, or as a Coulomb blockade device, has been illustrated. Further, such a device was experimentally realized and shown to operate in three different modes.

The gate-control ratio ($C_G/C_\Sigma$) for the simulated device structure (Fig. 5.9) is
Fig. 5.10. Functional simulation of a programmable SET device. (A) Device operation in Coulomb blockade (B) Electron tunneling across depletion barrier (C) Device operation as an Open (D) Device operation as a Short

Fig. 5.11. (A) Model of the programmable SET as a Coulomb blockade device (B) Coulomb oscillations of the programmable SET device using Monte-Carlo simulation.

0.4, where as it is 0.075 for the experimental device (Fig. 5.12), making these weak gate-control devices. Thus, this study helps us observe that practically realizable SET devices have weak gate-control.
5.4 Circuit Design Using SETs

5.4.1 Logic Gate Operation using SETs

In this sub-section, sub-200 mV digital logic operation at room temperature is illustrated and analyzed, using 2nm SETs with different gate-control ratios (as described in section 5.3.1). For complementary logic circuit operation, pull-down as well as pull-up devices are required. Because the Fermi level $E_F$ is assumed to lie in the middle of the energy gap ($E_0 + E_Q$), by symmetry, the SET device (Fig. 1.6A) behaves as a pull-up SET when negative drain and gate biases are applied. It is
assumed that, a SET device is $V_T$-adjusted such that the device operates at the $I_{\text{Valley}}$ when $V_{\text{Gate}} = 0$ and $V_{\text{Drain}} = V_{\text{CC}}$ (this assumption allows the SET device to operate with the best possible On/Off current ratio). Based on this assumption for SET operation, the $I_D$-$V_G$ curves of a pull-up SET and a pull-down SET with 2nm quantum-dot diameter, gate-control ratio = 0.7 and operating voltage $V_{\text{CC}} = 200 \text{ mV}$ are shown in Fig. 5.14A. The corresponding $I_D$-$V_D$ curves for the same are shown in Fig. 5.14B. The $I_D$-$V_G$ and $I_D$-$V_D$ characteristics for the 2nm SET device, with quantized energy levels, are obtained using discrete energy-level Monte-Carlo simulation in SIMON.

The Single Electron Transistor $I_D$-$V_G$ characteristics are populated into a VerilogA look-up-table model, using which circuit simulations shown in Fig 5.15 are carried out. The Voltage-Transfer Characteristics (VTC) of an inverter with good gate-control SET devices ($C_G/C_\Sigma = 0.7$) can be plotted for different $V_{\text{CC}}$ as shown in Fig. 5.15A. Fig. 5.15B shows the VTC of an inverter with weak gate-control SET devices ($C_G/C_\Sigma = 0.4$). As described in section 5.3.1), SET devices with poor gate-control have degraded SS, and cannot turn On properly when equal gate and drain voltages are applied. The consequence of this is clearly visible in Fig. 5.15B where the inverters have VTC curves with slope $< -1$, and hence cannot function as logic gates. A 2-input nand ring-oscillator circuit described in [87] is used to compare the Energy-Delay performance of SET-based digital logic with that of
5.4.2 Sense Amplifier-Based Design using SETs

The use of pass-transistor logic stacks with sense amplifiers was proposed previously [88, 89] in order to continue reducing energy when the circuit energy consumption becomes dominated by leakage energy at low-$V_{CC}$. As discussed in section 5.4.1, the energy consumption of complementary digital logic using SETs is leakage-energy dominated, thus motivating a similar approach using Sense-Amplifiers for SET-based logic. Furthermore, from the case-study in section 5.3.2, it is difficult to physically realize nanoscale devices with good gate-control, and it
is shown in section 5.3.1 that SET devices with poor gate-control require external amplification to drive subsequent logic stages. This is another reason to consider a Sense-Amplifier-based circuit design for SETs. Thus, we propose using SETs to implement Binary Decision Diagrams (BDDs) [90] of logic functions, followed by a differential Sense-Amplifier in order to reduce leakage and to provide amplification to drive the next logic level.

Similar to the concept in [88], the BDD logic stack does not consume any leakage energy during switching, since there are no paths-to-ground in it. The schematic of 8-input \textit{xor} logic using complementary SET gates is shown in Fig. 5.17, compared
to a SET BDD-sense amplifier implementation shown in Fig. 5.18 ([89]). The sense amplifier used for sensing the differential output of the SET-BDD logic stack is a current-controlled latch sense amplifier, originally proposed in [91], and is simulated using 16nm CMOS [25]. The root of the SET-BDD stack is driven with a supply voltage ranging from 125-to-275 mV, and the sense amplifier is driven with a constant supply voltage of 300mV. The output of the sense amplifier is in the voltage range of 0-to-300mV, and is used to drive subsequent logic stages. Fig. 5.19 shows the Energy-Delay performance of the 8-input Xor logic gate implemented using CMOS gates, and BDD-logic with sense amplifier. From Fig. 5.19, it is clear that the CMOS logic gates become leakage energy dominated at low-$V_{\text{cc}}$, and BDD-logic with sense amplifier helps to continue reducing energy even at low $V_{\text{cc}}$. However, it is also clear that the amount of energy reduction obtained using SET-BDD logic with 16nm CMOS sense amplifier is diminished, because the energy consumption is dominated by sense amplifier leakage energy (Fig. 5.19). The Minimum Energy of 8-input Xor logic BDD-sense amplifier is, 96 aJ for 22nm CMOS, 79 aJ for 16nm CMOS, and 76 aJ for 2 nm weak gate-control SET with 16nm CMOS sense amplifier.
5.4.3 SRAM cells using SETs

In addition to logic, the programmable SET device shown in Fig. 5.9 & Fig. 5.12 can also be used for building single electron SRAM cells. A negative differential conductance (NDC)-block composed of series-connected single-electron quantum-dots, shown in Fig. 5.20, was proposed in [92]. Fig. 5.21 shows the room temperature Monte-Carlo (MC) simulation of the output characteristics of the NDC-block for increasing gate control ($C_G/C_\Sigma$) ratio. Fig. 5.21 shows that the experimental programmable device in Fig. 5.12 with a $C_G/C_\Sigma$ of 0.075 cannot be used to obtain a NDC region. However, increasing the gate control $C_G/C_\Sigma$ ratio results in the appearance of a NDC-region when forward sweeping the drain voltage ($V_{DS}$). Monte Carlo simulation shows a NDC-region appearing at T=300K using a $C_G$ of 0.6 aF and a $C_G/C_\Sigma$ of 10.

Fig. 5.22 shows a circuit schematic of a hybrid CMOS/SET SRAM constructed using programmable SET devices [92]. The access transistor is a classical quantum well FET and the NDC-block is composed of devices programmed to operate in Coulomb blockade. Fig. 5.23 shows the simulation of the static characteristics of the hybrid SRAM at a reference voltage ($V_{Ref}$) of 450 mV.
**Fig. 5.22.** Schematic of SRAM based on NDC blocks [92]

**Fig. 5.23.** Load-line plot for single-electron SRAM showing the operating reference VCC and the stable binary states

**Fig. 5.24.** Layout comparison of (A) 6T CMOS SRAM and (B) 3T hybrid-NDC SRAM

**Fig. 5.25.** Area, Power, Delay Figure-of-Merit comparison for 6T CMOS SRAM and 3T hybrid MuQFET SRAM

Fig. 5.24 shows a layout comparison of a 6T SRAM using Si CMOS and hybrid-NDC SRAM using the programmable wrap-gate SET devices. 30% reduction in area is observed due to the use of fewer transistors. Fig. 5.25 compares SRAM figures of merit showing 75x reduction in dynamic power for the SRAM due to lower $V_{CC}$ operation.
5.5 Conclusion

In this chapter, the challenge of realizing sub-150mV logic operation using Single Electron Transistors has been studied. Analytical equations for the electrostatics and the peak/valley currents of SETs, as well as an approximation for including quantization energy, are derived. A design space for room temperature operation of SETs, based on the physical dimensions and the electrostatic properties of the device is described. By studying this design space, SETs can be classified as those with good electrostatic gate-control (which are suitable for complementary logic gate design), and those with weak electrostatic gate-control (which require a sense amplifier-based design). In order to reduce leakage, and also to provide the necessary sense-amplification for SETs with weak gate-control, we propose using SETs to implement the BDDs of logic functions. The use of BDD logic for SETs results in significant energy benefits over complementary CMOS or complementary SET logic. Due to the energy inefficiency of CMOS sense-amplifiers operating at low-$V_{CC}$, there is a need for design and exploration of low-power sense amplifiers in order to continue energy reduction with technology scaling.

This work has been featured on cover page of the Journal of Low-Power Electronics [93]. Subsequent work involving the experimentally demonstrated programmable split-gate SET has been accepted for presentation at the 2011 IEEE International Electron Devices Meeting (IEDM), to be held in Washington, DC, December, 2011 [21].
Conclusion

In this dissertation, we have modeled emerging nanoscale devices such as Single Electron Transistors and Interband Tunneling Transistors. Using a combination of analytical and numerical calculations, and using mixed-mode simulations, we have identified how these emerging transistors can be used to design circuits and systems which can beat the energy-delay performance of those based on modern silicon CMOS.

In chapters 2 & 3, sub-300mV logic and SRAM operation are explored. In chapter 2, we propose a heterogeneous TFET-CMOS multi-core architecture and illustrate a transistor-to-processor abstraction used for studying such an architecture. Using benchmark applications we illustrate the benefits of the heterogeneous TFET-CMOS architecture over homogeneous CMOS or TFET architectures. In chapter 2, the impact of variability on the noise margin characteristics of TFET circuits was ignored. This aspect is addressed in chapter 3. We demonstrate that though TFETs are prone to variation impact, require special structures such as a Schmitt-Trigger feedback can be used to ensure reliable low-\(V_{CC}\) operation. In chapters 2 & 3, the p-channel TFET is assumed to deliver on-current comparable to that of n-channel TFET, with a good on/off-current ratio. In chapter 4, we provide a physics-based analysis of p-channel TFETs, and illustrate that it is a challenging device design problem. Further, in chapter 4, we also propose a special type of SRAM cell taking advantage of asymmetric electrical characteristics of n-channel and p-channel TFETs. In chapter 5, we explore sub-150mV logic operation. We illustrate a physics-based framework for analyzing SET devices and
propose Binary Decision Diagram-Sense Amplifier (BDD-SA) circuits to achieve energy-efficient operation using SETs.

6.1 Future Work

Based on the research done in this dissertation we have identified several venues of future research. The architectural study in chapter 2 can be extended to study more complex energy-partitioning schemes based on a TFET-CMOS heterogeneous multi-core architecture. The variability analysis in chapter 3 can be extended to analyze other types of memory such as TFET-based embedded DRAM. Based on the analysis in chapter 4, the realization of a p-channel TFET may be challenging and further research on special circuit architectures for low-\(V_{\text{CC}}\) TFET-circuit operation may be necessary. Based on the study in chapter 5 the realization of low-\(V_{\text{CC}}\) sense-amplifiers is critical to the realization of energy-efficient BDD-SA circuit architectures.
Appendix A

Derivation of Analytical Expressions for $I_{\text{Valley}}$ and $I_{\text{Peak}}$ in a Single Electron Transistor

An analytical approximation for $I_{\text{Valley}}$ is derived. At the valley point ($V_{\text{Gate}} = V_{\text{Drain}}/2$), the Free Energy changes ($\Delta F$) corresponding to the various barrier-tunneling events across the source/nanodot and drain/nanodot tunnel barriers are shown in equations (A.1)-(A.4) [35].

$$\Delta F_{\text{Source} \rightarrow \text{Dot}} = \frac{E_0}{2} - \frac{q}{C_{\Sigma}} \left( C_D V_{\text{Drain}} + C_G \frac{V_{\text{Drain}}}{2} \right)$$

$$= \frac{1}{2} \cdot (E_0 - q \cdot V_{\text{Drain}}) \tag{A.1}$$

$$\Delta F_{\text{Dot} \rightarrow \text{Source}} = -\frac{1}{2} \cdot (E_0 - q \cdot V_{\text{Drain}}) \tag{A.2}$$

$$\Delta F_{\text{Drain} \rightarrow \text{Dot}} = \frac{E_0}{2} - \frac{q}{C_{\Sigma}} \left( C_D V_{\text{Drain}} + C_G \frac{V_{\text{Drain}}}{2} \right) + q \cdot V_{\text{Drain}}$$

$$= \frac{1}{2} \cdot (E_0 + q \cdot V_{\text{Drain}}) \tag{A.3}$$

$$\Delta F_{\text{Dot} \rightarrow \text{Drain}} = -\frac{1}{2} \cdot (E_0 + q \cdot V_{\text{Drain}}) \tag{A.4}$$

For reasonably small drain voltages, $\Delta F_{\text{Source} \rightarrow \text{Dot}} \gg k.T$ holds true i.e. $E_0 - q \cdot V_{\text{Drain}} \gg 2k.T$ holds. Under this assumption, the tunneling rates ($\Gamma$) for the
barrier tunneling events can be approximated as shown in equations (A.5)- (A.8). The tunneling rates (Γ) are computed assuming a constant density of energy states in the contact and the nanodot as described in [35].

\[
\Gamma_{\Delta F} = \frac{\Delta F}{q^2.R_T.(e^{\Delta F/k.T} - 1)} \quad \text{and} \quad \Delta F_{\text{Source} \rightarrow \text{Dot}} >> k.T
\]

\[
\Gamma_{\text{Source} \rightarrow \text{Dot}} \approx \frac{1}{2q^2.R_T} \cdot (E_0 - q.V_{\text{Drain}}) \cdot e^{-(E_0 - q.V_{\text{Drain}})/2k.T} \quad (A.5)
\]

\[
\Gamma_{\text{Dot} \rightarrow \text{Source}} \approx \frac{1}{2q^2.R_T} \cdot (E_0 - q.V_{\text{Drain}}) \quad (A.6)
\]

\[
\Gamma_{\text{Drain} \rightarrow \text{Dot}} \approx \frac{1}{2q^2.R_T} \cdot (E_0 + q.V_{\text{Drain}}) \cdot e^{-(E_0 + q.V_{\text{Drain}})/2k.T} \quad (A.7)
\]

\[
\Gamma_{\text{Dot} \rightarrow \text{Drain}} \approx \frac{1}{2q^2.R_T} \cdot (E_0 + q.V_{\text{Drain}}) \quad (A.8)
\]

Using this approximation, the occupation probabilities of the nanodot for the charge states N=0 and N=1 can be approximated as shown in equations (A.9)-(A.12).

\[
\Gamma_{0 \rightarrow 1} = \Gamma_{\text{Drain} \rightarrow \text{Dot}} + \Gamma_{\text{Source} \rightarrow \text{Dot}} \approx \frac{1}{2q^2.R_T} \cdot (E_0 - q.V_{\text{Drain}}) \cdot e^{-(E_0 - q.V_{\text{Drain}})/2k.T} \quad (A.9)
\]

\[
\Gamma_{1 \rightarrow 0} = \Gamma_{\text{Dot} \rightarrow \text{Drain}} + \Gamma_{\text{Dot} \rightarrow \text{Source}} \approx \frac{E_0}{q^2.R_T} \quad (A.10)
\]

\[
Pr(0) = \frac{\Gamma_{1 \rightarrow 0}}{\Gamma_{1 \rightarrow 0} + \Gamma_{0 \rightarrow 1}} \approx 1 \quad (A.11)
\]

\[
Pr(1) = \frac{\Gamma_{0 \rightarrow 1}}{\Gamma_{1 \rightarrow 0} + \Gamma_{0 \rightarrow 1}} \approx \frac{1}{2E_0} \cdot (E_0 - q.V_{\text{Drain}}) \cdot e^{-(E_0 - q.V_{\text{Drain}})/2k.T} \quad (A.12)
\]

Using the approximations for the occupation probabilities, equations (A.11) and (A.12), and considering the tunneling rates across the source-side tunnel barrier (Γ_{\text{Source} \rightarrow \text{Dot}} and Γ_{\text{Dot} \rightarrow \text{Source}}), an analytical approximation for \(I_{\text{Valley}}\) (the expression is multiplied by 2 to account for the contribution of the tunneling rates Γ_{0 \rightarrow -1}/Γ_{-1 \rightarrow 0}) is derived as shown in equation (A.13). Similarly an analytical expression for \(I_{\text{Peak}}\) is derived shown in equation (A.14).
\[ I_{\text{valley}} = 2q[Pr(0)\Gamma_{\text{Source} \rightarrow \text{Dot}} - Pr(1)\Gamma_{\text{Dot} \rightarrow \text{Source}}] \approx \frac{1}{q.R_T} \frac{(E_0 - q.V_{\text{Drain}})(E_0 + q.V_{\text{Drain}})}{2E_0} e^{-E_0/2kT} e^{qV_{\text{Drain}}/2kT} \quad (A.13) \]

\[ I_{\text{Peak}} = \frac{1}{2R_T} \frac{V_{\text{Drain}}}{2} \quad (A.14) \]
Bibliography


(MuQFETs),” in Accepted in IEEE International Electron Devices Meeting (IEDM), dec 2011.


Vita

Vinay Saripalli

Vinay Saripalli obtained his Dual-Degree (Bachelors & Masters) in Computer Science and Engineering from the Indian Institute of Technology (IIT) Madras in the year 2005. He joined the Ph.D. program in the Department of Computer Science and Engineering at the Pennsylvania State University, University Park, in Fall 2007, following two years of professional experience in the software industry. He has worked with the Microsystems Design Laboratory (MDL) and the Nanoscale Devices and Circuits Laboratory (NDCL) at Penn State, which are leading groups in academia involved in research on emerging nanoelectronic devices for high-speed, low-power, logic and embedded memory applications.

His Ph.D. dissertation involved technology-driven circuit and architecture design for ultra-low power logic application. Part of his research involved modeling emerging nanoscale transistors, such as Single Electron Transistors (SETs), Interband Tunneling Transistors (TFETs) and Quantum Well MOSFETs, which were then applied to novel circuit and architecture design. He has published various IEEE and ACM conferences and journals based on his research.

Partial List of Publications


- V. Saripalli, D. K. Mohata, S. Mookerjea, S. Datta and V. Narayanan, “Loadless 4T SRAM cell based on Degenerately Doped Source (DDS) InGaAs Tunnel FETs”, 68th Device Research Conference (DRC), Notre Dame, In, June 2010
