Exploring the Nano-Scale Self-Heating Mechanisms in SOI/Bulk MOS Devices

by

Krishna Pradeep [krishna.pradeep@epfl.ch]

Master Thesis Report

submitted to
School of Engineering
École Polytechnique Fédérale de Lausanne (EPFL)

in partial fulfillment of the requirements
for the degree of

Master of Science (2015)
In
Electrical and Electronics Engineering

supervised by
Prof. Yusuf Leblebici [yusuf.leblebici@epfl.ch]
Can Baltaci [can.baltaci@epfl.ch]
Microelectronic Systems Laboratory - STI/LSM, EPFL

Friday 19th June, 2015
If I have seen further it is by standing on ye sholders of Giants.
— Isaac Newton

To my parents…
Acknowledgements

First, I would like to thank Professor Yusuf Leblebici for giving me the opportunity to realize my master thesis at LSM on such an interesting and challenging topic. I would also like to especially thank my supervisor Mr. Can Baltaci for sharing his valuable experience, constant help and support. In particular, I would like to thank him for constantly being there for me, for patiently hearing out my ideas and doubts and giving his valuable suggestions. I would also like to thank Dr. (MER) Jean-Michel Sallese, Group Leader and Camillo Stefanucci, Research Assistant, EDLAB, EPFL for sparing their valuable time to discuss some of the issues which cropped up during the course of my work and for their valuable suggestions.

I'm also thankful to all the PhD students, scientists and master students at LSM for providing such a nice and inviting working environment. I would like to especially thank Marios Barlas, a fellow master student with whom I was sharing my desk, for the many enlightening and enjoyable discussions we had. I would also like to thank him for being away at CMI for most of the time, thereby granting me the use of the entire desk.

Finally, very special thanks go to my family, my mother Sheela and my father Pradeep for their love and support not only during my studies but also during all my life.

Lausanne, 19 June 2015

Krishna Pradeep
Abstract

The self heating effects inside both bulk and SOI MOSFETs were explored using hydrodynamic simulation models. Both transient and steady state analysis was performed to better understand the phenomena. The temperature inside the device was found to vary significantly during the operation of the device. With the understanding obtained from basic simulations explored, some structural modifications to improve the thermal efficiency of the device were explored. Also the dependence of the self heating on the circuit parameters was explored in detail.

Key words: FDSOI, MOSFETs, self heating, hydrodynamic simulations, transient, Sentaurus TCAD
List of Figures

1.1 Historical trends .......................................................... 1


2.2 3D view of the ellipsoidal conduction band valleys of Silicon within the first Brillouin zone. The different types of electron-phonon scattering are also graphically represented in the image. Image courtesy: [6] .................. 4

3.1 The output characteristics of SOI MOSFET with different effective gate lengths simulated with different simulation approaches. Image courtesy: [45] .......... 14

5.1 Simulated structure and the parameters used .......................... 17

5.2 Lattice temperature and heat generation at y=-5nm, Vds=2.1V, Vgs=1.2V .... 18

5.3 Lattice temperature and heat generation at Vds=Vgs=1V .................. 18

5.4 Lattice temperature and heat generation at y=-5nm (inside channel) Vgs=1V . . 19

6.1 Simulated structure and the parameters used .......................... 22


6.3 The lattice temperature and the total heat generation inside the device at the middle of the channel (y=-5nm) and at applied gate voltage of 1.2V. The plots made at different applied drain voltages as given in the legend .......... 24

6.4 Heat generation inside the device for different drain voltages and gate voltage=1.2V ...................................................... 25

6.5 Temperature distribution inside the device for different drain voltages and gate voltage=1.2V ...................................................... 26

6.6 Variation of maximum of lattice temperature rise and the heat generation and also cost indexes temperature rise/drain current and heat generation/drain current with applied drain voltage. The gate voltage is fixed at 1.2V ............ 27

6.7 Variation of maximum of lattice temperature rise and the heat generation and also cost indexes temperature rise/drain current and heat generation/drain current with applied gate voltage. The drain voltage is fixed at 1.2V .......... 27
6.8 Variation of maximum of lattice temperature rise and the heat generation and also cost indexes temperature rise/drain current and heat generation/drain current with applied gate voltage. The drain voltage is fixed at 1.2V. Modified with thermally insulated gate and threshold voltage of ∼250mV.

6.9 The solid lines correspond to case in which the gate, source and drain contacts are insulating and the dotted lines are for when the gate is insulating, while source and drain contacts are ideal heat sinks at 300K.

6.10 Plotted for the cases of insulated source and ideal heat sink drain (solid) and insulated drain and ideal heat sink source (dotted). In both figs 6.9 & 6.10 the gate voltage is fixed at 1.2V and the plots are made at the middle of the channel, ie y=-5nm. The red shaded area corresponds the channel regions, the blue the source and green the drain region.

6.11 The heat generation in the device for all the cases studied for different applied drain voltages. The gate voltage is fixed at 1.2V. The plots are made at the middle of the channel, ie y=-5nm. The red shaded area corresponds the channel regions, the blue the source and green the drain region.

6.12 Plots of the lattice temperature inside the channel for different source/drain contact lengths. Also plotted the output characteristics of the device with different contact lengths.

6.13 Structure used.

6.14 Comparison of the two structures.

6.15 Electrical curves for the gate length variation.

6.16 Thermal results of gate length variation. The lattice temperature and heat generation inside the channel plotted at a drain and gate voltage of 1V and in the middle of the channel (y=-5nm). The third plots the average lattice temperature with the applied power at the drain. An the final plot summarises the variation of the lattice temperature with the gate length.

6.17 Lattice temperature and heat generation inside the channel (y=-5nm) for an applied bias of 1.2V at both gate and drain.

6.18 The variation of the cost factors with applied gate voltage. The constant drain voltage of 1.2V is used.

6.19 The variation of different measured parameters with the spacing.

7.1 The variation of the lattice temperature with the applied pulse at the gate. Both the maximum and the average value of the lattice temperature are plotted. A constant dc voltage of 1V is applied at the drain.

7.2 Varying gate length.

7.3 Simulation setup used and characteristics of the gate signal used.

7.4 Variation of maximum and average temperatures inside the device, with time.

7.5 Variation of the mid value of the final oscillation with duty cycle and period of the input signal.

7.6 Drain current variation with the time.
7.7 Simulation setup used and characteristics of the input signal used............. 42
7.8 CMOS inverter operation ................................................................. 42
7.9 Variation of the peak and valley temperatures of the final oscillations with the period of applied signal ................................................................. 43
7.10 Fitting of the variations with a non-linear function ................................ 43
7.11 Variation of the peak and valley temperatures of the final oscillations with the rise and fall times at output ................................................................. 44
7.12 Fitting of the variations with a polynomial function (peak) and linear function (valley). ................................................................. 45
7.13 Variations with the ambient temperature ........................................... 45
7.14 Fitting of the variations of temperature rise with a polynomial function . . 46
7.15 An extreme case, with huge output load ............................................. 46
7.16 Variation with spacing ................................................................. 47
7.17 Structures considered ................................................................. 48
7.18 Effect of surroundings ................................................................. 49

B.1 Heat generation inside the device for different gate voltages and drain voltage=1.2V ................................................................. 65
B.2 Temperature distribution inside the device for different gate voltages and drain voltage=1.2V ................................................................. 66

C.1 Variation with drain voltage for different BOX thicknesses ......................... 67
C.2 Variation with drain voltage for different BOX thicknesses ......................... 67
C.3 Variation with gate voltage for different BOX thicknesses ......................... 68
C.4 Variation with gate voltage for different BOX thicknesses ......................... 68

D.1 Temperature distribution inside the device during the time in which the pulse is applied at the gate. The time scale used is the same as in fig 7.1 .................. 69
D.2 Temperature distribution inside the device during the time in which the pulse is not applied at the gate. The time scale used is the same as in fig 7.1 .................. 70
List of Tables

6.1 Gate workfunction differences .................................................. 30
6.2 Resistances used ................................................................. 31
# Contents

<table>
<thead>
<tr>
<th>Acknowledgements</th>
<th>i</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>List of figures</td>
<td>iv</td>
</tr>
<tr>
<td>List of tables</td>
<td>vii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Motivation</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Thesis Outline</td>
<td>2</td>
</tr>
<tr>
<td>2 Self Heating Mechanisms in Semiconductor Devices</td>
<td>3</td>
</tr>
<tr>
<td>2.1 Phonon Dispersion Model</td>
<td>4</td>
</tr>
<tr>
<td>2.2 Heat Conduction</td>
<td>5</td>
</tr>
<tr>
<td>2.3 Heat Generation</td>
<td>6</td>
</tr>
<tr>
<td>2.3.1 Drift-Diffusion approach</td>
<td>7</td>
</tr>
<tr>
<td>2.3.2 Hydrodynamic approach</td>
<td>7</td>
</tr>
<tr>
<td>2.3.3 Monte Carlo approach</td>
<td>8</td>
</tr>
<tr>
<td>2.4 Previous Works</td>
<td>8</td>
</tr>
<tr>
<td>2.4.1 Measurements</td>
<td>10</td>
</tr>
<tr>
<td>3 Hydrodynamic Simulations: Theory</td>
<td>11</td>
</tr>
<tr>
<td>3.1 Boltzmann's Transport Equation (BTE)</td>
<td>11</td>
</tr>
<tr>
<td>3.2 Band Structure</td>
<td>12</td>
</tr>
<tr>
<td>3.3 Higher Order Transport Models</td>
<td>13</td>
</tr>
<tr>
<td>3.4 Known Concerns</td>
<td>13</td>
</tr>
<tr>
<td>4 Simulation Setup Used</td>
<td>15</td>
</tr>
<tr>
<td>4.1 Structure and Meshing</td>
<td>15</td>
</tr>
<tr>
<td>4.2 Physics Used</td>
<td>15</td>
</tr>
<tr>
<td>4.2.1 Quantization Effects</td>
<td>16</td>
</tr>
</tbody>
</table>
5 Bulk MOSFET: Quasistationary Simulations
5.1 Varying gate lengths ........................................ 17
5.2 20nm MOSFET .................................................. 18
5.3 Observations ..................................................... 19

6 Fully Depleted SOI MOSFETs: Quasistationary Simulations
6.1 25nm Gate FDSOI .............................................. 21
6.2 Thin Film Thermal Conductivity .............................. 22
   6.2.1 Silicon ..................................................... 22
   6.2.2 Buried Oxide .............................................. 23
6.3 Quasistationary Simulations .................................. 24
6.4 Thermal Boundary Conditions ................................ 28
6.5 Realistic Device ................................................. 30
6.6 Gate Length Variation .......................................... 32
6.7 BOX Thickness Variation ...................................... 34
6.8 Gate contact/Source(Drain) Contact Spacing Variation 35
6.9 Observations ..................................................... 36

7 Transient Hydrodynamic Simulations ......................... 37
7.1 Transient Behaviour of Self Heating in a Single Device
   7.1.1 Behaviour over a Single Cycle ............................ 37
   7.1.2 Averaged Thermal Behaviour over many Cycles ......... 39
   7.1.3 Remarks ..................................................... 41
7.2 Transient Behaviour of Self Heating in an Inverter ......... 42
   7.2.1 Varying the Clock Frequency ............................ 43
   7.2.2 Varying the Output Load ................................ 44
   7.2.3 Varying the ambient temperature ......................... 45
   7.2.4 An Extreme Case ......................................... 46
   7.2.5 Remarks ..................................................... 46
7.3 Gate contact/Source(Drain) Contact Spacing Variation .... 47
7.4 Effect of Surroundings in a Circuit ......................... 48
7.5 Observations ..................................................... 49

8 Conclusion and Future Work .................................... 51

APPENDICES .......................................................... 54

A Physics used in Sentaurus TCAD ............................... 55
A.1 Electrode Definitions .......................................... 55
   A.1.1 Ohmic Contacts ............................................ 55
   A.1.2 Contacts on Insulators .................................... 56
   A.1.3 Resistive Contacts ......................................... 56
A.2 Thermal Boundary Conditions ................................. 56
A.3 Carrier Transport .............................................. 57
Chapter 1

Introduction

1.1 Motivation

The IC technology has come a long way from its humble beginnings in the hands of Jack Kilby at Texas Instruments in 1958 [1]. The number of transistors in a typical IC has grown enormously, and the dimensions of the transistor have shrunk. In 2014, both Intel[2][3] and Samsung[4] demonstrated circuits in the 14nm transistor nodes. With the shrinking dimensions, the power dissipation has been identified as a crucial bottleneck. This is even more true as we are dealing with huge static power dissipation due to leakage, which could have been easily ignored in primitive technology nodes.

Most manufacturers with the latest technological nodes have started concentrating on handling the power dissipation. The Broadwell technology (14nm technology) by Intel is an example of the same, which concentrate on low power processor design. With the changing market trend towards wearable electronics and smartphones, the designers are often restricted by the available power source for their design, while concentrating on the form factor. This calls for two different approaches: better battery technology or efficient/power conscious IC design.
With this realization, there has been a lot of research being conducted in understanding the power dissipation at the lowest level, i.e. self heating inside a single transistor. There has been quite a lot of varied research, with each group adopting a different approach to study the effects with a lot of argument as to which is the best. The confusion is further increased by the fact that experimental measuring of self heating in a single device is still at very nascent stages. Most of the measurements report the average temperatures in the device, whereas theoretical groups have long identified the existence of local hot spots in the devices, which cause more harm than the average temperature. Most of the theoretical works use either hydrodynamic or Monte Carlo steady state approaches to study the phenomenon.

The scope of this work is in the same lines, but we intend to concentrate on hydrodynamic simulations, due to its relative computational simplicity. We try to understand the temporal behaviour of the self heating phenomenon by performing transient simulations. We also try to explore some possible methods to modify the device structure to make it more power efficient, in this work.

1.2 Thesis Outline

This report is organized in 8 chapters, elucidating the theoretical background of the work done, the simulation setup used and the various experiments performed with the conclusions formed.

In the 2nd chapter, we describe the mechanism of self heating in semiconductor devices. We also present an overview of the previous works attempted by different groups in this topic.

In chapter 3, we describe the hydrodynamic simulation approach used. We explain the theoretical background of this method, and discuss its strengths and shortcomings.

In chapter 4, we describe the simulation setup we used for our experiments. We describe the concerned physics and methods as used by the tool we used, Sentaurus TCAD.

In chapter 5, we start describing the experiments we performed. The first study was the steady state exploration of bulk MOSFETs, which is explained in this chapter.

In chapter 6, we describe the steady state analysis of FDSOI MOSFETs. We explain the different experiments performed and derived conclusions, including the structural modifications.

In chapter 7, we describe the results of the transient simulations elaborating on the various experiments performed, and trying to reason out the observations made.

In the last chapter, we conclude our results, and also suggest some further works, which could provide more clarifications on the topic.
Chapter 2

Self Heating Mechanisms in Semiconductor Devices

Every transistor heats up during its operation due to the interaction of the carriers with the lattice. The new device structures proposed to negate the short channel effects, while moving forward according to the Moore’s law, have introduced geometries and materials which further reduce the thermal conductivity in the devices and has enhanced the self heating. [6] The self heating of the device brings into question the reliability of the device during its operation. As Vaselika et. al. mentions in her paper [7], the main reliability concerns are

- Degradation of electrical characteristics like the on and off currents of the transistor
- Self heating induced breakdown of the device
- Mechanical stresses due to the expansion of different materials in the device, resulting in a mechanical breakdown.
- Electromigration in the interconnects.

![Energy transfer processes in Silicon](image)

Figure 2.1 – Energy transfer processes in Silicon. Image courtesy: [6]
During the operation of a transistor, the free carriers are accelerated by electric field produced due to the applied potentials, and gain energy. These energetic carriers, as they move across the device can undergo different collisions. It is expected that the carriers loose its energy only by collisions with the lattice vibration (phonons), whereas only its momentum will be modified due to the collisions with other carriers, impurity atoms, lattice imperfections or interfaces. This energy transfer to the lattice results in increase of the lattice temperature, which in return affects the carrier transport in the device. This heating up of the lattice is termed as Joule heating. The electrons with energy lower than 50meV, scatter mostly with acoustic phonons in Silicon, whereas those with energies greater than 50meV scatter via optical phonons. In a MOSFET, the electric fields being pretty high, the scattering mainly occurs through optical phonons. But the optical phonons have a much lower group velocity of $\sim 1000\text{m/s}$ when compared to acoustic phonons which have group velocity of $\sim 5000 - 9000\text{m/s}$ (They also have higher occupation number). So the optical phonons do not contribute much to the heat conduction in Silicon, but they decay into the faster acoustic phonons. This decay rate is much slower compared to the electron-phonon (in a NMOS) scattering rates (fig.2.1). This creates a phonon energy bottleneck, which results in accumulation of optical phonons inside the device, which in turn impedes the carrier transport by causing more carrier scattering. \[6\]

### 2.1 Phonon Dispersion Model

![Figure 2.2 – 3D view of the ellipsoidal conduction band valleys of Silicon within the first Brillouin zone. The different types of electron-phonon scattering are also graphically represented in the image. Image courtesy: \[6\]](image)

Bulk silicon is reported to have 6 different phonon dispersion modes. \[8\]

- Two transverse acoustic modes (TA1 and TA2)
- One longitudinal acoustic mode (LA)
- Two transverse optical modes (TO1 and TO2)
- One longitudinal optical mode (LO)

The different electron-phonon and phonon-phonon scatterings are supposed to include these 6 phonon modes.

As shown in fig.2.2, there are 6 energetically equivalent valleys in the first conduction band (X-valley) of Silicon. The second conduction band (L-valley) can be safely ignored, as it lies slightly more than 1eV above the first and most of the nanoscale devices used for low power applications will be used at 1V or below. It is for the same reason that we ignore impact ionization in our simulations. The electron-phonon scatterings can mainly be of two types: intervalley and intravalley scatterings. The intervalley scatterings occurs between different valleys in the conduction band. They can still be of two different types: g-type, when the scattering occurs between the valley on the same axis ($\langle 100 \rangle$ to $\langle -100 \rangle$) and f-type when the scattering occurs between different axes ($\langle 100 \rangle$ to $\langle 010 \rangle$). So there are 6 types of intervalley scatterings, 3 of each f- and g-type. Intravalley scattering occurs when the electrons scatter within the same valley. This usually involves only the acoustic phonons.

In Silicon, the different phonon modes are supposed to possess different group velocities and different relaxation times for their interactions with other phonons and electrons. As mentioned earlier, the difference in these parameters result in the energy bottleneck inside the device, which causes the formation of the hot spots. It has been already argued by different authors that the granular approach for the different phonon modes is very important to predict the temperature and position of the local hot spots with reasonable accuracy.[6] [8] [7] [9]

The electron-phonon scattering rates can be calculated based on Fermi’s Golden Rule, which formulates the transition probability between two eigenstates,

$$P(k,k') = \frac{2\pi}{\hbar} |M(k,k')|^2 \delta(E_k - E_{k'} \pm \hbar \omega_q) \quad (2.1)$$

where $M(k,k')$ is the matrix element which describes the scattering process due to the lattice vibrations and the $\delta$-function prevents the overlap of collisions in space/time. The upper and lower signs corresponds to the absorption and emission of a phonon. The total scattering rate out of a state $k$ is found by integrating eq.2.1 over all the final states $k'$,

$$\Gamma(k) = \frac{2\pi}{\hbar} |M(k)|^2 g_d(E_k \pm \hbar \omega_q) \quad (2.2)$$

where $g_d$ is the density of states. It is also possible to derive the rates separately for the intervalley and intravalley scatterings. [6]

### 2.2 Heat Conduction

In semiconductors, the heat conduction has two components: the electronic contribution in which the free carriers are responsible for the heat transport and lattice vibrations. Contrary
to metals, which have a lot of nearly-free carriers, the electronic contribution accounts to much lower than 1% of the total heat transport in semiconductors. So the thermal conductivity of a semiconductor can be formulated as, \[ \kappa_s = \frac{1}{3} C_s \bar{v} \Lambda \] \hspace{1cm} (2.3)

where \( C_s \) is the heat capacity per unit volume, \( \bar{v} \) is the average phonon velocity and \( \Lambda \) is the average phonon mean free path. Fourier law gives the classical continuum heat diffusion equation in a material. \cite{6}

\[ C_s \frac{\partial T}{\partial t} = \nabla \cdot (\kappa_s \nabla T) + Q'''' \] \hspace{1cm} (2.4)

where \( Q'''' \) is the heat generation rate per unit volume in the material.

The issue with this formulation is that it cannot take into account the granularity of heat conduction in semiconductors as explained in previous paragraphs. A more true picture of the phonon transport at very small length scale (shorter than the acoustic mean free path) can be expressed by the Boltzmann Transport Equation (BTE), \cite{6}

\[ \frac{\partial f}{\partial t} + v \cdot \nabla f = \left( \frac{\partial f}{\partial t} \right)_{\text{coll}} + \left( \frac{\partial f}{\partial t} \right)_{\text{g}} \] \hspace{1cm} (2.5)

where \( f(\mathbf{r}, \omega, t) \) is the phonon distribution function and \( v \) is the phonon velocity. The first term in the RHS, which is due to the phonon collisions, can be approximated using the relaxation time approximation as,

\[ \left( \frac{\partial f}{\partial t} \right)_{\text{coll}} = f_0 - f \frac{1}{\tau_{\text{ph}}} \] \hspace{1cm} (2.6)

where \( f_0 = 1/(\exp(h\omega/k_B T) - 1) \) is the equilibrium Plank distribution at temperature \( T \) and \( \tau_{\text{ph}} \) is the average phonon scattering time. The eq.2.5 can be integrated over the phonon frequency and density of states to get,

\[ \frac{\partial u}{\partial t} + v \cdot \nabla u = \frac{u_0 - u}{\tau_{\text{ph}}} + Q'''' \] \hspace{1cm} (2.7)

in terms of the phonon energy density, \( u \). \cite{6} The different simulation approaches uses different solutions of the BTE to solve the heat conduction in the device. They also use different approaches for identifying the heat generation term, \( Q'''' \).

### 2.3 Heat Generation

The exact identification and calculation of the heat generation term in the device is as important as the heat conduction model used in the device. Historically, the heat generated inside a material, through which a current, \( I \) is flowing due to the application of a potential \( V \) across the material, can be given as,

\[ Q = I \times V \] \hspace{1cm} (2.8)
This is a very basic approach and has many shortcomings. It fails to account for the voltage drop across contact resistances. Also in the case of quasi-ballistic devices where hot electrons might escape from the device before transferring the energy to the lattice, eq. 2.8 overestimates the heat generation inside the device. This equation is also insufficient to identify the local hot spots, as it only provides an estimate of the total heat generated. [6]

### 2.3.1 Drift-Diffusion approach

The Drift-Diffusion approach is the simplest one used to simulate semiconductor devices. In this approach, the current density inside the device is calculated based on the drift of the carriers due to the applied potential and the diffusion due to the carrier concentrations. In this approach, ignoring the contribution from the holes (minority carriers in a NMOS), the heat generated can be calculated as,

\[
Q'' = J \cdot E + (R - G)(E_g + 3k_B T) \tag{2.9}
\]

where \(J = q n v_e\), with \(n\) being the electron density and \(v_e\) the average electron velocity. The first term of eq. 2.9 gives the Joule heating in the device, which is usually positive. The second term gives the contribution from the carrier generation and recombination. [6]

This formula is known to have two major shortcomings. [6]

- This method does not take into account the microscopic non-locality of the phonon emission near the peak of the electric field. The electrons tend to lose the energy gained (maximum of which corresponds to peak in the electric field) by collisions with the lattice, but only in multiples of phonon energy. This would result in a displacement of a few electron mean free path lengths between the maximum electric field point and the complete transfer of the energy to the lattice. While this spatial spread can be easily ignored for longer devices, at the recent nanoscale devices this approximation is no longer true.

- Also this formulation fails to differentiate between the different phonon types and ignores the interactions between them.

### 2.3.2 Hydrodynamic approach

The heating rate is calculated in the more sophisticated hydrodynamic approach using, [6]

\[
Q'' = \frac{3}{2} k_B \frac{n(T_e - T_L)}{\tau_{e-L}} + (R - G) \left[ E_g + \frac{3}{2} k_B (T_e + T_L) \right] \tag{2.10}
\]

where \(T_e\) is the electron temperature, \(T_L\) is the lattice temperature and \(\tau_{e-L}\) is the energy relaxation time. Once again, here the holes are ignored are assumed to be in thermal equilibrium with the lattice. This method has better resolution of the non-locality of the phonon emission near the peak electric field, but has problems due to the simplifications of using only a single carrier temperature and energy relaxation times. Also as in eq. 2.9 this approach ignores the differentiation between the different types of the phonons emitted.
2.3.3 Monte Carlo approach

It has been widely agreed that only a simulation approach which takes into consideration all the scattering events and distinguishes between the different types of phonons can fully describe the heat generation inside the device with accuracy. The Monte Carlo approach seems to fit the bill quite well, as it takes into consideration all the carriers in the device and keeps track of their collisions. It can be modified to take into consideration the disparity between the different phonon modes and different authors have incorporated these interactions to varying degrees. The heat generation rate in Monte Carlo approach can be expressed as, \[ Q'' = \frac{1}{t_{sim}} \Sigma (\hbar \omega_{ems} - \hbar \omega_{abs}) \] (2.11)

where \( t_{sim} \) is the simulation time.

2.4 Previous Works

Some of the earlier approaches tried modifying the hydrodynamic simulation to include the interaction between the electrons and phonons by solving the BTE for the phonons also. But most of these authors failed to differentiate between the different phonon modes.

Fiegna et al.\[11][12\] calibrated a drift-diffusion (DD) model at different temperatures by comparison with a full-band self-consistent MC simulator for a 3-D electron gas with corrections to the electrostatic potential in order to include the effect of carrier quantization on the spatial distribution of the inversion charge. They modified the mobility parameters in the DD model to match the isothermal transfer and drain characteristics obtained from two simulators and studied the self heating effects in different SOI MOS structures including FINFETs. On the other hand Sadi et al.\[13\] tried to iteratively couple a 2D Monte Carlo simulator with a 2D steady state heat diffusion equation (HDE) solver to study electrothermal effects in a Si/SiGe MODFET. But both these approaches are insufficient in identifying the thermal localization in the devices, due to the inherent issue with the steady state heat diffusion formulation.

The group of Vasileska et al.\[7][14][9\], tried to solve the coupled electron-optical phonons-acoustic phonons-heat bath problem using energy balance equations derived from the Boltzmann equations. This was coupled with a Monte Carlo electric simulation to study the self heating effects in SOI devices. Even though their phonon transport model was not very exhaustive, and fairly simple, they did report some interesting results from their simulations. They argued that the quasi-ballistic transport of the electrons in the short channel devices, results in very less lattice heating in the channel, which improves the thermal degradation of these devices when compared to longer ones \[7][9\]. Rodriguez et al.\[15\], also made similar conclusion regarding the average temperature rise in the channel, which they measured during the high frequency operation of the SOI MOSFET.
Some of the earlier developments can be grouped as "gray" which regarded the phonon transport with a single group velocity and relaxation time\[16\] and "semi-gray" which considered two different mutually interacting phonon modes: reservoir mode, responsible for the heat generation and included the longitudinal optical and acoustic phonons, and propagation mode including the transverse optical and acoustic phonons and responsible for the heat conduction \[17\][18]. The results from these kinds of approaches will strongly depend on the choice of the relaxation time\[8\].

The later works tried to concentrate on the inclusion of the differentiation between the different types of the phonons. Narumanchi et al.\[19\] came up with a BTE model, which incorporated ballistic term for the acoustic modes, but ignored the optical modes. Wang\[20\] developed a method for computing the scattering rates for the different phonon modes from the perturbation theory. He showed that the optical phonons played a prominent role in determining the hotspot temperature. At this point, it might be interesting to point out the disparity among different authors about the dominant phonon mode responsible for determining the hotspot temperature. In contrast to Wang’s argument, Pop\[6\][21][22] claimed that it is the longitudinal optical modes which were dominant. Whereas Nghiêm et al.\[23\] argues that it is longitudinal acoustic and transverse optical modes. The Monte Carlo simulation developed by Pop\[6\], used an analytic, non-parabolic electron energy band to model electron transport while using an isotropic, analytic phonon dispersion model for electron-phonon scattering, distinguished between the different phonon modes, but they did not concentrate much on the phonon transport. \[8\]

Sinha et al.\[24\] developed a split-flux model for non-equilibrium phonon BTE model, which calculated the phonon transport based on a phonon generation spectrum obtained from a Monte Carlo simulation. Rowlette et al.\[25\] further modified this model by fully coupling split-flux model with the Monte Carlo simulation.

One of the main drawbacks of the method developed by Wang\[20\], was that the computation is very expensive. A simpler anisotropic relaxation time phonon BTE model was developed by Ni\[26\], where the relaxation time is a function of the wave vector. The group in University of Illinois at Urbana-Champaign coupled their in-house improved version of the full band Monte Carlo simulation tool\[27\] with the anisotropic relaxation time model\[28\]. The group reported a temperature rise of 46.5K in the drain edge of the channel in a 50nm bulk MOSFET\[8\]. They also concluded that there is still significant self heating effects in the channel of the short channel devices\[28\] which is in contrast to earlier findings by Vasileska et al.\[7][9\]. On the other hand Nghiêm et al.\[23\], who developed yet another Monte Carlo approach, concluded that the self heating effects have limited impact on the DC characteristics of a 20nm double gate MOSFET.

A recent work by Zhang et al.\[29\] derived a new expression for the thermal conductivity of the phonons starting from the Holland’s model. They modified the model by including the effects
of phonon scattering of boundary reflections and bound electrons using the Matthiessen’s rule. The temperature distribution was calculated using a finite element method (FEM). They claimed that their results match those of Vasileska et al.\[7\]\[9\]. Misawa et al.\[30\] reported the development of a Monte Carlo approach with different time scales for the electron and phonon transport, in which the group velocities of the phonons were calculated from the dispersion curves reported by Azuhata et al.\[31\]. Muscato et al.\[32\] came up with an improved formulation to calculate the heat generation due to the electron-phonon scattering in a Monte Carlo approach,

\[
\langle H \rangle(x) = \frac{1}{N_{\text{obs}}} \sum_{j=1}^{N_{\text{obs}}} \left[ \frac{n}{N_{\rho}} \sum_{i=1}^{6} n_{\epsilon} G(\epsilon(k_i(t_j))) \right],
\]

\[
G(\epsilon) = \sum_{i=1}^{6} \hbar \omega_i [\lambda_i^+ - \lambda_i^-] \quad (2.12)
\]

\(N_{\text{obs}}\) is the total number of observation times \(t_j\). They argued that the variance of the eq\[2.12\] is lower than of eq\[2.11\] and hence it provides a better estimate of the heat generated. Rhyner and Luisier\[33\] used an atomistic quantum transport simulator coupling electron and phonon transport to study self heating in Si nanowire transistors, whereas Burenkov et al.\[34\] used molecular dynamics simulations to study self heating in nano-scale SOI MOSFETs.

### 2.4.1 Measurements

In spite of considerable work being done in the theoretical development of the self heating effects in devices and the characterization of the local hot spots in the device, there is not much practical verification of the same. The practical verification is plagued, at the foremost, by the inability to achieve the kind of resolution required to identify the local hot spots in a device. Most of the practical work concentrate on the extraction of the average channel temperature and identifying the effects on the performance of the device. The measurements are usually performed using two different approaches: Time domain characterization which usually includes applying a pulsed IV setups and frequency domain characterization which uses AC conductance techniques.\[35\].
Chapter 3

Hydrodynamic Simulations: Theory

For the problem in hand about the self heating in nanoscale devices, it has been identified that a drift-diffusion approach will not be sufficient. The assumptions underlying the drift-diffusion model loses their validity when the dimensions of the device are smaller than the mean free path of electrons and phonons. Even from the thermal point of view, it has already been illustrated that drift-diffusion model is insufficient. On the other extreme of the scale, there has been quite a lot of recent advancements with fully coupled Monte Carlo algorithms solving the BTE, which are ideally suited to solve these kinds of problems. But the huge computational requirements and the noisy results imbibes a huge limitation to its application for device design on a regular basis. The Monte Carlo approach provides one of the best solutions for creating a benchmark, but for other purposes one needs to find a compromise between the computational complexity and the extent of physics simulated. A hydrodynamic simulation provides a very good balance between the two.[36][37][38][39][40]. Owing to the complexities of Monte Carlo approach, Muscato and Stefano [41] claims that obtaining a hydrodynamic model from BTE equations with an ideal closure for the higher order moments is a good engineering oriented solution to study the electro-thermal behaviour of the devices.

3.1 Boltzmann’s Transport Equation (BTE)

The transport in a semiconductor can be be generally modelled using the BTE, which is a semiclassical kinetic equation [42], given by,

\[
\frac{\partial f}{\partial t} + \mathbf{u} \cdot \nabla_{\mathbf{r}} f + \frac{\mathbf{F}}{\hbar} \cdot \nabla_{\mathbf{k}} f = C[f]
\]  

(3.1)

where \( f(\mathbf{k}, \mathbf{r}, t) \) is the carrier distribution in the six-dimensional space and \( C[f] \) is the rate of change of \( f \) due to collisions. The group velocity \( \mathbf{u} \) is given by,

\[
\mathbf{u}(\mathbf{k}, \mathbf{r}) = \frac{1}{\hbar} \nabla_{\mathbf{k}} \varepsilon(\mathbf{k}, \mathbf{r})
\]  

(3.2)

where \( \varepsilon \) is the carrier kinetic energy. The inverse mass tensor is given by,

\[
\hat{m}^{-1}(\mathbf{k}, \mathbf{r}) = \frac{1}{\hbar} \nabla_{\mathbf{k}} \otimes \mathbf{u}(\mathbf{k}, \mathbf{r}) = \frac{1}{\hbar^2} \nabla_{\mathbf{k}} \otimes \nabla_{\mathbf{k}} \varepsilon(\mathbf{k}, \mathbf{r})
\]  

(3.3)
where \( \otimes \) represent the tensor product. \( \mathbf{F} \) is the force exerted on electrons in the presence of electric and magnetic fields and for inhomogeneous material is given by,

\[
\mathbf{F}(\mathbf{k}, \mathbf{r}) = -\nabla_{\mathbf{r}} (E_{c,0}(\mathbf{r}) + \varepsilon(\mathbf{k}, \mathbf{r})) - q(\mathbf{E}(\mathbf{r}) + \mathbf{u}(\mathbf{k}, \mathbf{r}) \times \mathbf{B}(\mathbf{r}))
\]

(3.4)

The first two position dependent terms account for changes in the bottom of the conduction band edge \( E_{c,0} \) and the shape of the band structure. If we ignore the magnetic field and assume homogeneous material,

\[
\mathbf{F}(\mathbf{r}) = - q \mathbf{E}(\mathbf{r})
\]

(3.5)

To solve this equation numerically by discretization of the differential and integral operators is a very computationally intensive task. A common simplification is to use some moments of the distribution function (different orders), which is the method employed in hydrodynamic simulations. Different authors chose differently the moments and accordingly they derived the closure expressions. A moment is obtained by multiplying the distribution function with a suitable weight function, \( \Phi = \Phi(\mathbf{k}) \) and integrating over the \( \mathbf{k} \) space [42],

\[
\langle \Phi \rangle = \frac{1}{4\pi^3} \int \Phi f d^3k
\]

(3.6)

### 3.2 Band Structure

The main concern in any transport model is how the band structure is modelled in the material. One of the common assumption is an isotropic approximation for the band structure— with the dispersion relation [42],

\[
\gamma(\varepsilon) = \frac{\hbar^2 k^2}{2m^*}
\]

(3.7)

Assuming a parabolic relation between carrier energy and momentum, [42],

\[
\gamma(\varepsilon) = \varepsilon
\]

(3.8)

But this is valid only when the energy is close to the band minimum. A more appropriate expression is the non-parabolic relation, [42]

\[
\gamma(\varepsilon) = \varepsilon(1 + \alpha \varepsilon)
\]

(3.9)

with \( \alpha \) being the non-parabolicity factor. This relation would give the following relation between the momentum and velocity [42],

\[
\mathbf{u} = \frac{\hbar \mathbf{k}}{m^*} \frac{1}{1 + 2 \alpha \varepsilon} = \frac{\hbar \mathbf{k}}{m^*} \sum_{i=0}^{\infty} (2 \alpha \varepsilon)^i
\]

(3.10)

As can be observed in eq (3.10), the average velocity now contains infinite number of higher order terms, which might not be negligible. This makes it almost non-tractable. A simpler approximation was introduced to solve this issue [42],

\[
\gamma(\varepsilon) = x \varepsilon^y
\]

(3.11)
with fitted parameters \( x \) and \( y \). With this approximation, the density of states can be expressed as

\[
g(\varepsilon) = \frac{8\pi}{\hbar^3} (2m^* x)^{3/2} y^{(3/2)y-1} = g_0 \varepsilon^\lambda
\]  

(3.12)

For the evaluation of the moments which require an integration over the entire energy range, a value less than 1 is required for \( \lambda \) to fit the low energy region, where \( f(\varepsilon) \) is maximum [42].

\section*{3.3 Higher Order Transport Models}

One needs to transform the BTE into a macroscopic transport model to apply it for device simulations. This is done using the moments and corresponding closure relations. The review by Grasser \textit{et al.} [42] gives a nice overview of the different derivations attempted.

Two of the earliest derivations were done by Stratton and by Bløtekjær. Stratton [43] used a microscopic relaxation time approximation for the collision operator and split the distribution function into odd and even parts. Bløtekjær [44] on the other hand took the moments of the BTE with weight functions one, \( \hbar k \) and \( \varepsilon \) without imposing any assumptions on the distribution function. He used a macroscopic relaxation time approximation for the collision operators. Many other authors also derived the macroscopic transport models similar to the one by Bløtekjær, but which relied on an Anstaz for the distribution functions [42].

There has been quite a few works which tried to extend the hydrodynamic formulation for degenerate semiconductors, to include non-parabolicity extensions, multiple bands, band-splitting and electrothermal applications. Interested readers can refer to the review by Grasser \textit{et al.} [42] for further details.

\section*{3.4 Known Concerns}

There are some approximations made in this approach, which brings into question of the validity of the approach. [42]

- One of the critical issue is about the closure. Usually the moments will generate more unknown variables than the equation available at disposal. This calls for the creation of some closure relation, which in effects invokes some relations between the different order moments, and in effect brings in some approximations.

- Approximation of tensor quantities appearing in the formulation with scalar, such as carrier mass and carrier temperature.

- In the approximation of the average carrier energy, the contribution of the drift energy is usually neglected. It has been shown that in some cases, this approximation may not be true as the convective energy can often be comparable to the thermal values.
• The approximation of relaxation times is another great concern. The hydrodynamic approach usually uses values derived from homogeneous field measurements or Monte Carlo simulations.

• Models based on Bløtekjær’s method have been reported to show spurious velocity overshoot. It is argued that this is due to the hysteresis in mobility or due to non-parabolicity effects or due to the closure used.

• With decreasing critical dimensions of the device, the carrier transport is becoming intrinsically nonstationary. For $L_c=100-50\text{nm}$, this implies quasiballistic regime, where a universal hydrodynamic model is rather difficult. So with these dimensions, it becomes even more important to verify the simulation results with a competent Monte Carlo or other methods.

![Figure 3.1 – The output characteristics of SOI MOSFET with different effective gate lengths simulated with different simulation approaches. Image courtesy: [45]](image)

Even with the above mentioned validity concerns, hydrodynamic approach provides a good compromise between the computational complexity of the simulations and the completeness of physics. As it was discussed before, the computational complexity of an electro-thermal Monte Carlo simulator is too huge which makes optimization of device design impossible with that approach. So the best approach will always be to use the simpler hydrodynamic simulations while validating the results with some other method.
Chapter 4

Simulation Setup Used

We used TCAD simulation tools provided by Sentaurus for performing our simulations. The workbench tool provided by Sentaurus makes an ideal environment to organize your different experiments and run them in an efficient manner.

4.1 Structure and Meshing

The device structure was created using the structure editor, which provides a graphical/script based user interface to design the structure using simple geometries. It also provides the functionality to perform the doping in the structure and to define required meshing strategies. An analytical doping profile with Gaussian distribution was used for the drain and source regions, whereas a constant doping profile was applied for the channel and substrate regions. Advanced doping mechanisms like halos and gradient doping in the channel were ignored, as we were more interested in studying the self heating in a simple structure. The meshing is done using the Sentaurus mesh tool, with default parameters. The tool generates Delaunay meshes containing triangle for 2D structures and tetrahedrons for 3D structures.

4.2 Physics Used

The simulations were performed in the Sentaurus device tool, which is in effect a numerical solver. The predominant method of solving used in our simulation is the hydrodynamic simulation model with the temperature equation implemented by means of energy balance equations. Both the carrier temperatures were calculated in addition to the lattice temperature.

We activated the effective bandgap narrowing model to include the effects of temperature and doping on the bandgap of Silicon. The Philip's unified mobility model proposed by Klaassen was used with saturation of mobility at higher fields. Only Shockley-Read-Hall recombination was considered in our simulations. Direct band recombination is not expected to occur in an indirect band gap semiconductor like Silicon. Similarly, the Auger
recombination was also ignored, as the doping densities in the channel was not very high. Impact ionization was also ignored as the devices considered are mainly to work in the sub 1V regime, where impact ionization is not expected to occur. The physics as used by Sentaurus is explained in more details in appendix A.

4.2.1 Quantization Effects

As we are dealing with nanoscale devices, it makes sense to include quantization models in the simulations to take into account the quantum effects. But in our simulations, we opted to ignore these. Two considerations were used to reason out our decision:

- It has been reported that the quantum effects only affects the current density in the device, but makes hardly any effect in the temperature profile in the device[37]. This was also verified with pilot simulations in our case.

- The inclusion of the quantization models makes the simulations much more computationally complex. We observed a notable increase in the simulation times and also had problems with the convergence even with simple simulations.

So for obtaining a good enough approximation of the phenomena occurring in the device, within the constrained time, we decided to avoid these in our simulations.
Chapter 5

Bulk MOSFET: Quasistationary Simulations

The MOSFET is the elemental building block for almost all electronics circuits. In this chapter, we would attempt to investigate the self heating in a MOSFET. As mentioned before, a hydrodynamic simulation approach was adopted. 2D quasistationary simulations were performed for different gate lengths. The structure used is given in fig. 5.1. The boundary conditions used were ideal heat sinks fixed at 300K for the drain and source, while all the other boundaries were assumed to be thermally insulating. The contact on the left side was taken as the source and the right side was taken as the drain.

**Figure 5.1 – Simulated structure and the parameters used**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gate oxide thickness</td>
<td>5nm</td>
</tr>
<tr>
<td>Substrate thickness</td>
<td>500nm</td>
</tr>
<tr>
<td>Source/Drain contact length</td>
<td>100nm</td>
</tr>
<tr>
<td>Source/Drain peak doping</td>
<td>$1 \times 10^{20}$</td>
</tr>
<tr>
<td>Channel doping</td>
<td>$1 \times 10^{17}$</td>
</tr>
<tr>
<td>Polysilicon gate</td>
<td>$1 \times 10^{20}$</td>
</tr>
</tbody>
</table>

5.1 Varying gate lengths

The metallurgical gate length was varied between 20nm and 200nm. The remaining parameters for the device were kept constant.
As can be observed in fig. 5.2, the maximum lattice temperatures increases with decreasing gate length. But the temperature rise is not very big. The majority of the heating occurs in the drain edge of the channel, as predicted by previous work [6], but there is considerable heating also occurring in the channel. This could be due to the fact that the devices considered here are considerably longer compared to the electron-phonon scattering length. As explained in [8], the electrons which gain energy in the channel is expected to travel a few scattering lengths before losing the energy to the phonons. This explains the almost negligible heating in the channel in the case of ultra short devices (5-10nm). But this is not expected to happen in the longer devices simulated in this chapter, which show some significant heating also in the channel.

![Figure 5.2 – Lattice temperature and heat generation at y=-5nm. Vds=2.1V, Vgs=1.2V](image1)

5.2 20nm MOSFET

![Figure 5.3 – Lattice temperature and heat generation at Vds=Vgs=1V](image2)

The temperatures observed inside the device (a little over 3K) were not very high even for a relatively small gate length of 20nm. The majority of the temperature rise occurs in the drain edge of the channel. Similarly the peak of the heat generation was also observed in the drain.
end of the channel. The trend can be observed in more detail in fig. 5.4 for different drain voltages. The gate voltage was kept at 1V.

![Graphs](image)

(a) Lattice temperature distribution  
(b) Total heat generation

Figure 5.4 – Lattice temperature and heat generation at y=-5nm (inside channel) Vgs=1V

### 5.3 Observations

The results observed in our simulations were similar to those reported by Muscato and Stefano [41]. In their work, they reported the simulation of a 1D n+/n/n+ diode, which is supposed to mimic the behaviour of MOSFET channel. They reported a temperature rise of $\sim 2K$ in the channel and heat generation in the device had similar distribution as in our case. They were assuming a channel length of $\sim 100nm$. On the other hand, other authors who used complex electro thermal Monte Carlo simulators with detailed phonon transport formulations [8] [28] [19] [25] [7], reported much higher values of temperatures in the channel. They argued that the difference was due to the fact that a hydrodynamic model ignores the various interactions between the different types of phonon, which play a big role in the heat transport in the device. Ni et al. [8] reported a temperature rise of $\sim 46.5K$ in a 50nm bulk MOSFET using their electrothermal Monte Carlo simulator which uses an anisotropic relaxation time approximation for the phonon interactions.
Chapter 6

Fully Depleted SOI MOSFETs: 
Quasistationary Simulations

The SOI (Silicon on Insulator) technology has been introduced as a solution for many of the issues faced by conventional bulk devices. Instead of the isolation of the devices using reverse biased pn junctions, SOI technology uses dielectrics like silicon dioxide to isolate the devices from each other, which decreases source and drain capacitances and improves the speed of the devices. Also it reduces the leakage currents in the device and prevent latch up issues. Two types of structures have been proposed for SOI MOSFETs: fully depleted (FD-SOI) and partially depleted (PDSOI). In the case of FDSOI, the body of the transistor is thin enough that the entire body region gets depleted. In contrast, only a part of the body region is depleted in the case of PDSOI. PDSOI is known to have some issues due to the floating body, which accumulates charges and creates anomalous behaviour like the presence of the kink in the output characteristics, which can be solved using FDSOI. FDSOI is also known to possess sharper subthreshold curves, and the DIBL can be regulated by changing the ratio between the body thickness and the BOX thickness.

In spite of the benefits of FDSOI, the structure is not very favourable from a thermal point of view, due to the much lower thermal conductivity of the dielectric (100 times lower for $SiO_2$ compared to Silicon). This makes these devices more susceptible to self heating effects when compared to bulk devices. So it becomes quite important to study the self heating effects in SOI devices in more detail, which we will be attempting in this chapter.

6.1 25nm Gate FDSOI

For the sake of our study, we mainly used a gate length of 25nm, with a structure as shown in fig 6.1. For initial simulations, all the contacts were assumed to be thermally insulated except for the substrate and gate contact which were assumed to be ideal heat sinks at 300K. The contact on the left side is assumed to the source and that on the right side is the drain in our simulations.
Thin Film Thermal Conductivity

(a) Structure simulated

(b) Parameters used

Figure 6.1 – Simulated structure and the parameters used

The maximum depletion region in the channel is given by,

\[ W_{d, \text{max}} = \sqrt{\frac{4 \epsilon_S \psi_B}{q N_A}} \text{, where } \psi_B = \frac{k_B T}{q} \ln \left( \frac{N_A}{n_i} \right) \]  \hspace{1cm} (6.1)

Substituting the values of doping used in eq. 6.1 one gets the maximum depletion length to \( \sim 100 \text{nm} \), which is much larger than our body thickness.

6.2 Thin Film Thermal Conductivity

It is known that when the thickness of the material is smaller than the phonon mean free path in that material, the thermal conductivity of the material drops drastically. This is mainly thought to be due to the increased phonon-boundary scattering.

6.2.1 Silicon

The phonon mean free path is \( \sim 300 \text{nm} \) in Silicon. For a thickness of \( 20 \text{nm} \), the thermal conductivity is almost an order of magnitude lower than that of bulk [7]. In a Monte Carlo simulator which solves the BTE, this effect is already taken into consideration as explained by Asheghi et al. [47], but not in a hydrodynamic approach. So we used a modified version of the approach of Sondheimer [48].

This approach assumes that the phonon boundary scatterings are purely diffusive. For a semiconductor film of thickness \( t \), and with the \( z \)-axis being perpendicular to the plane of the film located between \( z = 0 \) to \( z = t \), the thermal conductivity as function of \( z \) is given by [7],

\[ \kappa(z) = \kappa_0(T) \int_0^{\pi/2} \sin^3 \theta \left\{ 1 - \exp \left( -\frac{t}{2 \lambda(T) \cos \theta} \right) \cosh \left( \frac{t - 2z}{2 \lambda(T) \cos \theta} \right) \right\} d\theta \]  \hspace{1cm} (6.2)
where $\lambda(T)$ is the mean free path given by,

$$\lambda(T) = \lambda_0 \left( \frac{300}{T} \right) \text{nm} \quad (6.3)$$

with room temperature mean free path, $\lambda_0 = 290 \text{nm}$. On the other hand, Sentaurus uses the following expression

$$\kappa_0(T) = \frac{1}{a + \frac{bT}{c} + \frac{cT^2}{d}} \text{W cm}^{-1} \text{K}^{-1} \quad (6.4)$$

where $a = 0.03 \text{cmK W}^{-1}$, $b = 1.56 \times 10^{-3} \text{cmW}^{-1}$ and $c = 1.65 \times 10^{-6} \text{cmK}^{-1} \text{W}^{-1}$. Sentaurus does not provide the possibility of defining a distance dependent thermal conductivity inside the material. To overcome this issue, we integrated eq.6.2 over the thickness of the film,

$$\kappa_{av} = \int_0^t \kappa(z)dz = \frac{\kappa_0(T)}{t} \left[ \frac{2t}{3} - \lambda(T) \int_0^{\pi/2} \sin^3 \theta \cos \theta \left( 1 - \exp \left( -\frac{t}{\lambda(T) \cos \theta} \right) \right) d\theta \right] \quad (6.5)$$

For a thickness of 10nm, as in our device and at a temperature of 300K,

$$\kappa_{av} = 0.06574317\kappa_0(T) \quad (6.6)$$

where $\kappa_0(T)$ is given by eq.6.4. The values of the parameters $a$, $b$ and $c$ were modified in the parameter file according to eq.6.6 in the parameter file for the thin body. The default equations were used in other regions. A modified value of $0.102 \text{W cm}^{-1} \text{K}^{-1}$ was found for the 10nm thin Silicon body.

Figure 6.2 – Silicon thin film thermal conductivity dependence on the temperature and film thickness. Image courtesy:[7]

### 6.2.2 Buried Oxide

At this point it might be interesting to study if a similar thickness dependent thermal conductivity variation can be expected for the BOX. We were unable to find much information in this regard, but according to Goodson et al.[49], the frequency independent phonon mean free path length in amorphous silicon dioxide is 4.94Å. If we use a similar reasoning as before, no thickness dependent thermal conductivity variation is expected unless the thickness of the oxide film is lower than this value.
There is considerable variation in the reported values for the thermal conductivities of silicon dioxide. Goodson et al.\textsuperscript{50} reported a value of $1.4 \text{W m}^{-1} \text{K}^{-1}$ for BOX thickness more than 300nm, fabricated by separation by implantation of oxygen (SIMOX), which is also the default value used in Sentaurus. Tenbroek et al.\textsuperscript{51} measured values of 0.66 and 0.82 $\text{W m}^{-1} \text{K}^{-1}$ for 320nm and 420nm BOX layers fabricated by wafer bonding and SIMOX methods, respectively. More recently, Dong et al.\textsuperscript{52} and He et al.\textsuperscript{53} reported a values of $\sim 0.92 \text{W m}^{-1} \text{K}^{-1}$ and $\sim 1.06 \text{W m}^{-1} \text{K}^{-1}$ respectively. Both papers converge in reporting that the measured values stays constant for all the BOX thickness measured (10nm to 55nm). This is in line with our earlier reasoning.

### 6.3 Quasistationary Simulations

The simplest simulations to study the effects of self heating in the device is a quasistationary simulation, which assumes steady state conditions. We attempted two different types of quasistationary sweeps:

(a) Lattice temperature

(b) Total heat generation inside the device

![Graphs showing lattice temperature and total heat generation](image)

Figure 6.3 – The lattice temperature and the total heat generation inside the device at the middle of the channel ($y=-5\text{nm}$) and at applied gate voltage of 1.2V. The plots made at different applied drain voltages as given in the legend

- Drain voltage was swept between 0V and 2.1V, while keeping the gate voltage fixed at 1.2V
- Gate voltage was swept between 0V and 2.1V, while keeping the drain voltage fixed at 1.2V

It is interesting to study the distribution of the lattice temperature (fig\textsuperscript{6.5}) and heat generation (fig\textsuperscript{6.4}) in the device. The 2D plots of the distribution of lattice temperature and heat generation for the gate voltage variation can be found in the appendix\textsuperscript{9}.

The observed results are very close to the ones reported by Vasileska et al.\textsuperscript{7} \textsuperscript{14} \textsuperscript{9} and Etes-sam et al.\textsuperscript{39} for similar device. As can be observed in the previous plots, the temperature rise happens mainly in the drain side. There could be two reasoning for this observation.
From fig. 6.3b, we could identify that the majority of the heat generation seems to occur in the drain edge of the channel which implies an almost ballistic transport in the channel with hardly any electron-phonon scattering. This is even more true for lower applied voltages. This is similar to the argument introduced by Vasileska et al.\cite{7,14,9} to conclude lower self heating effects for short channel devices. But in our case, for higher applied voltages, significant heating seems to be occurring in the channel also which is in contradiction to the argument of Vasileska et al. But such a behaviour was suggested by Mohamed et al.\cite{28} who reported considerable self heating effects in the channel for short channel devices, due to the spreading back of the heat generated in the drain region into the channel.

![Figure 6.4 – Heat generation inside the device for different drain voltages and gate voltage=1.2V](image)

At the same time, the temperature rise in the channel region is very low, compared to that in drain region. This could probably be due to the unrealistic thermal boundary condition of an ideal sink assumed for the gate. In an LSI circuit the gate of an individual transistor will be buried under many layers of metal and dielectric, and also input might be originating from a higher temperature. The work by Vasileska et al.\cite{7} concentrated mainly on the lattice temperature distributions in the device rather than on the heat generation. Also they were using boundary conditions similar to ours, which could probably be the reason for their conclusions, which were later argued to be faulty by Mohamed et al.\cite{28}.
Quasistationary Simulations

Figure 6.5 – Temperature distribution inside the device for different drain voltages and gate voltage=1.2V
Quasistationary Simulations

Figure 6.6 – Variation of maximum of lattice temperature rise and the heat generation and also cost indexes temperature rise/drain current and heat generation/drain current with applied drain voltage. The gate voltage is fixed at 1.2V

Fig. 6.6 makes the trend of the temperature rise and the heat generation in the device with the applied voltages clearer. We defined the cost factors “Temperature rise/drain current” and “Heat generation/drain current” and observed their variations with the applied drain voltages. The more interesting trend is that of heat generation. The cost factor has an almost linear relation with the applied drain voltage but there seems to be some residual saturation effect still observed in the trend. This would mean that the heat generation is having an effect of some other parameter in addition to the drain current.

Figure 6.7 – Variation of maximum of lattice temperature rise and the heat generation and also cost indexes temperature rise/drain current and heat generation/drain current with applied gate voltage. The drain voltage is fixed at 1.2V

The trends are even more interesting to observe with respect to the applied gate voltages. As can be observed in figs. 6.7a & 6.7b, the cost indexes temperature rise/drain current and heat generation/drain current are both much higher for lower applied gate voltages. This would imply that the cost of the leakage current or subthreshold currents in the device in terms of the self heating is quite enormous when compared to the on state operation of the device.

There are two main concerns with the above explanation. The first is the negative threshold voltage of our device, which results in considerable drain current even with a gate voltage
of 0V. Another concern is the false thermal boundary conditions that we had adopted for this experiment. With this in mind a second simulation was attempted in which the gate contact work function was adjusted to get a threshold voltage of \( \sim 250 \text{mV} \) and with insulating boundary condition for the gate (fig.6.8). It can be observed that the cost in terms of heat generated is higher for a positive threshold voltage. This emphasizes how damaging it is for the device in terms of the heat generated, to have a big leakage current. On the other hand, the trend of cost in terms of lattice temperature is quite different. It is almost identical for both off and on states of the device. The cost has a minima at around a gate voltage of \( \sim 1 \text{V} \). This is a very valuable information, as most of these nanoscale devices are often operated at sub 1V regime. According to our results, it also seems that the cost of operating the device in terms of the temperature rise is also minimum at this regime.

![Figure 6.8](image_url)  
(a) Maximum heat generation/drain current  
(b) Maximum temperature rise/drain current

Figure 6.8 – Variation of maximum of lattice temperature rise and the heat generation and also cost indexes temperature rise/drain current and heat generation/drain current with applied gate voltage. The drain voltage is fixed at 1.2V. Modified with thermally insulated gate and threshold voltage of \( \sim 250 \text{mV} \).

### 6.4 Thermal Boundary Conditions

The thermal boundary conditions need to be fixed properly considering how the device will be situated in a LSI circuit. For our simulations, we always assumed the substrate to be an ideal heat sink with a temperature of 300K. This is a reasonable approach as in a circuit this will be the contact whose temperature can be easily controlled externally. As for the gate, drain and source, it is much more difficult to apply any external cooling. For reasoning given earlier, one can assume an insulating thermal boundary condition for the gate. For the source and drain contacts, the conditions are similar to that of the gate, but these are more easily accessible in a circuit. For this purpose, we try different boundary conditions for these. As for the sides of the devices, an insulating thermal boundary condition was assumed. In an LSI circuit, the device is most probably going to be surrounded by other devices generating a similar heat profile. So it can be safely assumed that no heat transfer will occur across the neighboring devices, on an average.
Thermal Boundary Conditions

Figure 6.9 – The solid lines correspond to case in which the gate, source and drain contacts are insulating and the dotted lines are for when the gate is insulating, while source and drain contacts are ideal heat sinks at 300K.

Figure 6.10 – Plotted for the cases of insulated source and ideal heat sink drain (solid) and insulated drain and ideal heat sink source (dotted). In both figs.6.9 & 6.10, the gate voltage is fixed at 1.2V and the plots are made at the middle of the channel, i.e., y = -5nm. The red shaded area corresponds the channel regions, the blue the source and green the drain region.

For the two extreme cases of ideal heat sink and thermal insulation for source/drain, the lattice temperatures observed in the device differs significantly (fig.6.9a). The maximum temperature rise is 50K and 350K respectively for ideal heat sink and insulated boundary conditions for an applied voltage of 2.1V at drain and 1.2V at gate. Such enormous temperatures can be reasoned out because there is hardly any heat transfer from the device in the second case. The only path for the heat to be removed from the device is through the substrate contact which is separated from the channel by the thick BOX, which has very small thermal conductivity. Whereas in terms of the heat generated inside the device, the difference between the two cases is not as drastic (fig.6.9b). The heat generation rate is higher in the first case. This is expected due to the higher current inside the device made possible by the lower lattice temperatures.

The next experiment considered was to study the effect of making either the source or the drain contact insulating while keeping the other as an ideal heat sink (fig.6.10). The insul-
ing boundary condition resulted in higher temperatures in that region. Also it was observed that having an insulated drain contact was more detrimental for the device from a thermal perspective. Due to the presence of the peak of the heat generation at the drain edge of the channel, it is more thermally efficient, if one is able to achieve more cooling for the drain contact. Fig 6.11 summarizes the heat generation in all cases studied. It can be observed that the heat generation is more or less similar in all cases. The biggest deviation is found for the case with insulated gate, drain and source contact, for which a relatively lower heat generation was observed.

![Figure 6.11 – The heat generation in the device for all the cases studied for different applied drain voltages. The gate voltage is fixed at 1.2V. The plots are made at the middle of the channel, i.e., y=-5nm. The red shaded area corresponds the channel regions, the blue the source and green the drain region.](image)

### 6.5 Realistic Device

Before proceeding further, it was decided to make some optimization to the device to make it more realistic and at the same time simpler to simulate. At first, the gate was modified to give a positive threshold voltage for the NMOS by adjusting the workfunction difference between the gate metal and silicon in the simulation. We adjusted this parameter to get a threshold voltage of $\sim250mV$ for NMOS and $\sim-250mV$ for PMOS. The workfunction difference used are given in table 6.1. In a circuit, the substrate is possibly going to much thicker. To illustrate the same we assumed a substrate thickness of 5$\mu$m. Also the metal contacts of source drain are a few metal layers long. To take into account the same, we introduced 1$\mu$m long copper contacts at source and drain. It was observed that the temperature rise inside the device increased with the length of the contact used (fig 6.12a).

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NMOS</td>
<td>-0.12eV</td>
</tr>
<tr>
<td>PMOS</td>
<td>0.1eV</td>
</tr>
</tbody>
</table>

Table 6.1 – Gate workfunction differences

To keep the computational complexity in check, it was decided to emulate the effect of the
longer contacts and substrate using thermal and electric resistances. The thermal resistances can be calculated as,

$$R_{th}[KW^{-1}cm^2] = \frac{l[cm]}{k[Wcm^{-1}K^{-1}]} \quad (6.7)$$

where \( l \) is the length of the material and \( k \) is the thermal conductivity of the material which for Silicon is \( 1.3 W cm^{-1} K^{-1} \) and for copper \( 4.01 W cm^{-1} K^{-1} \)\textsuperscript{[54]}\textsuperscript{[55]} . The electrical resistance, which is expressed as the distributed resistance in Sentaurus for the contact can be calculated as,

$$R_{el}[\Omega cm^2] = \rho[\Omega cm] \times l[cm] \quad (6.8)$$

where \( l \) is the length of the material and \( \rho \) is the electrical resistivity which for copper is \( 1.6 \times 10^{-8} \Omega m \) \textsuperscript{[55]}.

![Figure 6.12](image)

(a) Lattice Temperature
(b) Output characteristics

Figure 6.12 – Plots of the lattice temperature inside the channel for different source/drain contact lengths. Also plotted the output characteristics of the device with different contact lengths.

The calculated values were used as a starting point for the optimization and later adjusted to get similar electrical and thermal behaviour for both the (fig.6.13). We were able to get very identical electrical behaviour in both cases. But in the case of the thermal behavior, we were able to match the profiles in the channel, but under the source and drain contact this was not possible. But we decided to go ahead with this approximation as most the important phenomenon is occurring in the channel, where we have similar behaviour. The optimized values of the resistances used are given in table 6.2. For the substrate, a thickness of 1nm below the BOX was assumed to avoid placing the contact directly on the oxide.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source/Drain thermal resistance</strong></td>
<td>( 2.1 \times 10^{-3} cm^2 K/W )</td>
</tr>
<tr>
<td><strong>Substrate thermal resistance</strong></td>
<td>( 4 \times 10^{-4} cm^2 K/W )</td>
</tr>
<tr>
<td><strong>Source/Drain distributed electrical resistance</strong></td>
<td>( 60 \times 10^{-11} \Omega cm^2 )</td>
</tr>
</tbody>
</table>

Table 6.2 – Resistances used


### 6.6 Gate Length Variation

In this section we tried to vary the physical gate length ($L_g$), while keeping the remaining physical dimensions constant. The gate length was varied between 25nm and 250nm. Both the electrical and thermal behaviour of the devices were studied for the different gate lengths.

The improvement of the drain current for shorter gate length is well documented, and the same can be observed in our results (fig. 6.15). Also threshold voltage shift with decreasing gate length is also observed. The curves observed also include the self heating effects in the device. The lattice temperature rise in the device is more for the shorter device. This is in contradiction to the observation of Vasileska et al.\[7\], but in line with the observations of Mohamed et al.\[28\]. Also the heat generation inside the device is also lower in the longer devices which could be due to the decrease in the drain current. The lower temperatures could be due to the lower heat generated in the device. It was also found that the peak of the heat
Gate Length Variation

...generation moves more into the drain for shorter devices (fig 6.16b with the x axis normalized with Lg), which can be attributed to the quasi ballistic transport in shorter devices. The slope of curves in fig 6.16c gives the net thermal resistance of the device. From this we can observe that the net thermal resistance is more or less the same for all devices. This is in agreement with the trends observed in the current and temperature in the device.

(a) Output characteristics, gate voltage=1V  
(b) Transfer characteristics, drain voltage=1V

Figure 6.15 – Electrical curves for the gate length variation

(a) Lattice temperature  
(b) Heat generation

(c) Average lattice temperature v/s power applied  
(d) Variation of maximum lattice temperature with gate length

Figure 6.16 – Thermal results of gate length variation. The lattice temperature and heat generation inside the channel plotted at a drain and gate voltage of 1V and in the middle of the channel (y=-5nm). The third plots the average lattice temperature with the applied power at the drain. An the final plot sumarises the variation of the lattice temperature with the gate length.
6.7 BOX Thickness Variation

Tuning the BOX thickness has often been regarded as a manner to compensate the self heating effects in FDSOI. As we had discussed earlier in this report, the thermal conductivity of silicon dioxide stays almost constant for varying film thickness. In this experiment we varied the BOX thickness from 50nm to 10nm and observed the self heating behaviour of the devices.

![Graphs showing lattice temperature and total heat generation](image)

Figure 6.17 – Lattice temperature and heat generation inside the channel (y=-5nm) for an applied bias of 1.2V at both gate and drain.

As can be observed from fig.6.17, the temperatures inside the device is lower for thinner BOX. This is expected, as a thinner BOX will allow better cooling of the device from the substrate contact. But the improvement in the temperature is only 10K. At the same time, the heat generation in the device is mostly the same, with a tiny decrease with decreasing BOX thickness.

The variation of the cost factors in terms of temperature rise and heat generation per drain current is similar for all BOX thicknesses (fig.6.18). The cost in terms of heat generation increases with decreasing BOX thickness, in the subthreshold regime. But in the on state the cost is mostly constant with varying BOX thickness. On the other hand the cost in terms of temperature rise displayed opposite trend with the BOX thickness. The cost mostly decreased with decreasing BOX thickness. In this case, the effect is more prominent in the on state, where the thinner BOX showed a much lower cost in terms of the lattice temperature rise. Similar trends were observed for the cost indexes as a function of applied drain voltages. But the effect of varying BOX thickness was much less prominent, especially for the temperature rise.

From these results, it can be understood that reducing the BOX thickness is indeed helpful in lowering the temperature inside the device, but the cost in terms of the heat generation per drain current goes higher with decreasing BOX thickness, especially in the subthreshold regime. So the BOX thickness of the actual device can be tuned by taking into consideration the operating regime of the device under consideration, to gain the minimum self heating.
Gate contact/Source(Drain) Contact Spacing Variation

Spacing between the gate and the source/drain contacts is a parameter which is known to affect the behaviour of the device. It has been predicted that bringing the two contacts close to each other is good both from electrical and thermal perspectives. From electrical perspective, this will lower the source/drain parasitic resistance, whereas from thermal point of view, bringing them closer will help cooling the channel better. But in an actual device one is limited by technological resolution in determining this spacing. Also bringing the contacts closer is known to increase the parasitic capacitances in the device, which might hamper the high frequency operation of the device. With these considerations, we decided to try vary the spacing between 5nm and 80nm while keeping the other parameters constant. The total volume of the device simulated was kept constant in all the simulations. [6][21].

As expected, the lattice temperature decreased with decreasing spacing (fig[6.19d]). The saturated drain current and the calculated threshold voltage also seem to decrease with increasing spacing (figs[6.19b] & [6.19c]). For better understanding, we performed some isothermal simulations using drift-diffusion (DD) models. The results of both DD and hydrodynamic (HD) simulations are shown in fig[6.19c]. The variation in the threshold voltage is not very prominent for the isothermal case as it is for the thermal case. So the threshold voltage shift is caused by the increased temperatures. Similar lowering of the threshold voltage was reported by Groeseneken et al.[56], who measured the threshold voltage of SOI MOSFETs at different temperatures. The lower saturated drain current can be attributed to the reduction in the carrier transport due to the higher temperatures. Esseni et al.[57] reports a dependence of the order -1.4 with temperature for the mobility of thin SOI films.

The trend of the maximum heat generated in the device (fig[6.19e]) cannot be understood as easily as the other trends. For better understanding, we integrated the total heat generated inside the device and compared it with the power applied to the device in terms of applied costs.

Figure 6.18 – The variation of the cost factors with applied gate voltage. The constant drain voltage of 1.2V is used.

6.8 Gate contact/Source(Drain) Contact Spacing Variation

(a) Lattice temperature rise/drain current
(b) Total heat generation/drain current
Observations

36

drain voltage times the drain current. The results are shown in fig. 6.19. It was observed that for longer spacing, the two curves approach each other. But when the contacts are closer, the heat generated in the device is much lower than power applied to the device. This could be attributed to the escape of hot electrons through the contacts before it transfers the entire energy to the lattice by collisions. It might be required to perform simulations including heat dissipation in the contacts to verify this hypothesis.

(a) Parameter varied

(b) Saturated drain current

(c) Threshold voltage

(d) Maximum lattice temperature

(e) Maximum heat generation

(f) Power spent in device

Figure 6.19 – The variation of different measured parameters with the spacing.

6.9 Observations

From the quasistationary studies, we found that the self heating increases with shorter gate lengths and that having a thinner BOX and making the gate and source/drain contacts helped reduce the temperature rise inside the device.
Chapter 7

Transient Hydrodynamic Simulations

This chapter is going to describe the transient hydrodynamic simulations attempted using a FDSOI NMOS. The objective of these experiments was to study how the self heating inside the device manifests during the high speed operation of the device in some digital circuit, and to observe the temporal manifestation of self heating in the device. The optimized 25nm device structure introduced in the previous chapter was used for the simulations.

7.1 Transient Behaviour of Self Heating in a Single Device

In a circuit, the device hardly ever stays on for a long period of time. The device keeps switching during its operation. It would be interesting to study how the self heating inside the device manifests during the switching of the device.

7.1.1 Behaviour over a Single Cycle

The first experiment performed was to observe the heating inside the device as a function of time. For this purpose, the drain was connected to a dc supply of 1V, and a single pulse signal was applied at the gate, with an amplitude of 1V, pulse width of 10ns and rise and fall times of 10ps each. The simulation was run for 21ns, so as to give equal time for the heating and cooling of the device, also accounting for the initial 1ns delay.

As can be observed from fig[7.1] the maximum lattice temperature rises very fast with applied pulse at the gate, but it slows down at higher temperatures. The slowing down can be reasoned if one considers the heat conduction away from the device, which increases with the temperature difference between the device and its surroundings. If we compare these values with the results shown in fig[6.14] we find that the saturated value is similar to the values obtained in the quasi-stationary simulation with similar biasing. The time taken for the maximum lattice temperature to reach its saturated value is a lethargic 4-5ns, whereas the it reached approximately 90% of its saturated value in less than 1ns, which is pretty fast. In comparison, the average lattice temperature rises at a much slower pace, and saturates at approximately half of the maximum lattice temperature.
Figure 7.1 – The variation of the lattice temperature with the applied pulse at the gate. Both the maximum and the average value of the lattice temperature are plotted. A constant dc voltage of 1V is applied at the drain.

Another important observation from this experiment, is the fact that the device cools down to the ambient temperature quite fast. The cooling down is slightly slower than the heating up, but both the average and maximum lattice temperature falls back to ambient temperature in around 5-6 ns.

It is quite imperative to note that the device heats up to its maximum temperature in a very short time, but it also cools down in a similarly short time. So, in a digital operation with a clock of period $\geq 10$ns and a duty cycle of 50%, the device should not display any net heating. Even when there is no expected net heating inside the device, the performance of the device can be expected to degrade due to self heating of the device during the short time that the device conducts.

Observing the change of the spatial distribution of temperature inside the device helps in identifying the region where any temperature driven material breakdown can be expected to occur. It was observed that the temperature distribution is very similar to the ones observed for quasistationary simulations. Once the pulse is turned off, the temperature lowers much faster in the silicon compared to the BOX which is also expected due to the much lower thermal conductivity of $SiO_2$ compared to that of Silicon. The 2D plots of the spatial distribution of the temperature can be found in appendix D.

Varying Gate Length

In this section we discuss the effect of varying the gate length, which we discussed earlier in previous chapter, on the transient self heating behaviour of the device. As before, we varied the metallurgical gate length between 25nm and 250nm and applied the same input signal at the gate. The results are shown in fig. 7.2.
It was found that rise and fall of the temperature becomes slower with increasing gate length. It is more prominent in the case of the average lattice temperatures, where the time taken to reach steady state are much higher for longer devices. Also the cooling of the device, once the gate signal is removed, takes a longer time for the longer devices. More thermal energy can be expected to be stored in the longer devices, which takes longer time to discharge. The argument can be verified if we compare the temperature rise in terms of the maximum and the average lattice temperatures. The maximum lattice temperature takes almost similar time to saturate whereas the average lattice temperature takes longer to saturate in the longer devices, due to the bigger area in these devices.

### 7.1.2 Averaged Thermal Behaviour over many Cycles

Next, we tried to apply a pulse train with a higher frequency at the gate, so as to prevent the device from cooling back to ambient temperature within each clock cycle. The aim of this experiment was to study the cumulative effect due to the residual temperature in each clock cycle. The simulation was run for a longer time and the results were analyzed.

![Simulation setup used and characteristics of the gate signal used](image)

From fig. 7.4, it can be observed that the lattice temperature inside the device does not rise to the values seen earlier (fig. 7.1). As the width of the applied pulse is smaller than the time required...
for the saturation of the temperature, there seems to be a net increase of the temperature over time. The net rise seems to be much faster at the start of the applied input, but it slows down later to settle down to oscillations between two fixed temperatures. For confirmation, two different simulations were performed: starting from an already ON state and an already OFF state. In both cases, the steady state achieved was observed to be highly perturbed about an average temperature. Similar trends were observed for both maximum and average lattice temperatures.

![Graph](image)

Figure 7.4 – Variation of maximum and average temperatures inside the device, with time.

On comparison with fig. 7.1, it can be observed that the final oscillations seems to be occurring around a temperature which is approximately half of the saturated temperature values found before. We tried to derive a formula to predict the mid value for the final oscillations from the values obtained from quasistationary simulations. Let's assume \( \Delta T_{sta} \) to be the value of lattice temperature rise obtained from quasistationary simulations, then the mid value of the final oscillations can be defined as,

\[
T_{mid} = 300 + \Delta T_{sta} \ast DutyCycle
\]  

(7.1)

![Graph](image)

(a) Fitting with the formula. The lines are plotted according to eqn. 7.1, while the dots represent data obtained from simulation

(b) Variation of the mid value of the final oscillation for the same duty cycle (50%) with different period.

Figure 7.5 – Variation of the mid value of the final oscillation with duty cycle and period of the input signal.
To verify this assumption, a few experiments were performed with different duty cycles and with same duty cycle but with different period, and the results were compared with eqn [7.1]. As can be observed from fig [7.5], the average lattice temperature follows the proposed formula quite well, while the maximum lattice temperature has some slight variations from the proposed relation.

The drain current is a good indication of the performance of the device, in terms of the load it can drive. As can be observed from fig [7.6], the drain current depreciates with the self heating. The average drain current when the device is ON is ~1.4516mA/µm. It is found to decrease slightly with the increase in net temperature (fig [7.6a]). During each half period, in which the device stays ON, the drain current decreases following the variation in lattice temperature (fig [7.6b]). A variation of ~0.025 mA/µm is observed during each clock cycle. This shows that the drain current follows the variation of the temperature in the device very closely.

![Fig 7.6 - Drain current variation with the time](image)

(a) The general trend of the drain current in comparison to that of maximum lattice temperature  
(b) The variation of drain current in each cycle in comparison to that of maximum lattice temperature

### 7.1.3 Remarks

The temporal behaviour of the self heating in a FDSOI MOSFET was analyzed for the first time, to the best of our knowledge, using transient hydrodynamic simulations. The 25nm device simulated took approximately 5ns for lattice temperature to achieve steady state values. This information is very important, as one could assume that if the operation frequency of the device is lower than the ~0.2GHz, no net heating can be expected in the device. Whereas for higher frequencies, a net temperature was observed in the device, in which case the lattice temperature was observed to settle into oscillation. We were able to derive a formula to predict the mid value for the oscillations based on the results from quasistationary simulations.
7.2 Transient Behaviour of Self Heating in an Inverter

The next objective was to study the self-heating occurring in the device when it is operating in an actual circuit. A CMOS inverter was chosen for this purpose as it is supposed to have maximum activity.

![CMOS inverter](image)

<table>
<thead>
<tr>
<th>Period</th>
<th>200ps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rise and fall times</td>
<td>25ps</td>
</tr>
<tr>
<td>ON voltage</td>
<td>1V</td>
</tr>
</tbody>
</table>

Figure 7.7 – Simulation setup used and characteristics of the input signal used

![CMOS inverter operation](image)

(a) Variation of maximum lattice temperature in both the FETs  
(b) Variation of maximum lattice temperature

Figure 7.8 – CMOS inverter operation

It was observed that the temperature rise was more prominent in the NMOS than in PMOS. The fraction of the time during which the devices are actually conducting in the circuit is very small (« 25% of the period of the input signal) and is dependent mainly on the rise and fall times at the input and output. This implies that the device gets heated up for a much shorter period than it has for cooling down, which is a favourable situation from the thermal perspective of the device. As can be observed from fig. [7.8] the net temperature in the devices behaved much similar to the previous case. The final oscillations were at a much lower temperature here. For average lattice temperature, it was hardly half a degree higher than the ambient temperature.
We were interested in studying how the temperature behaved with the output load of the inverter, the clock frequency of the applied input and the ambient temperature. To perform these simulations using the entire inverter circuit with both the FETs would be quite computationally demanding. So it was decided to concentrate on the NMOS alone. For this purpose, we applied the same input at the gate, but at the drain we applied another pulse train, whose rise and fall times were adjusted to match the ones in an inverter.

### 7.2.1 Varying the Clock Frequency

The clock frequency was varied for pulses applied both at gate and drain. The rise and fall times were kept constant at 10ps at the gate and 25ps at the drain. A small time difference of 5ps was assumed between the gate and drain pulses to account for the gate propagation delay. The ambient temperature was set to be 300K.

The lattice temperature reduces exponentially with increasing period of the signal. A significant temperature rise is observed in the GHz range of the applied signal. On increasing the period, the time during which the device conducts remains constant, whereas the time available for the device to cool down increases. This would mean that decreasing the clock...
frequency beyond a certain value will have no further effect in the net heating of the device, as device already cools down to ambient temperature in each clock cycle. The lattice temperature then keeps oscillating between the ambient temperature and the temperature that rises during the short period in which the device conducts. This can be observed from fig 7.9 where the valley temperature saturates at 300K (ambient temperature) for lower frequencies and the peak saturates around 306K. The increase in the lattice temperatures for higher frequencies is due to the additive increase of the lattice temperature, caused by the incomplete cooling of the device in each cycle. In case of a device which is thermally well insulated from the surroundings inside a circuit, a very high frequency operation can result in very temperatures inside the device.

7.2.2 Varying the Output Load

The purpose of this experiment was to study the effect of the output load of the inverter on self heating. For this purpose, we varied the rise and fall times of the drain pulse, while keeping the remaining parameters constant (period of 200ps, rise and fall times at the gate at 10ps and a delay of 5ps between the pulses at input and output). For better modelling of the real case, we also performed the same simulation using a resistive load inverter by varying output capacitive load. The results of both the simulations are shown in fig 7.11.

![Graph](image)

(a) For adjusted signals at gate and drain

(b) For resistive load inverter

Figure 7.11 – Variation of the peak and valley temperatures of the final oscillations with the rise and fall times at output

The lattice temperature increases with increasing rise and fall times at the output, ie. drain. The peak values of the final oscillations seems to follow a 2nd order polynomial relation with the rise and fall times, whereas the valley values seems to have a linear dependence. For the resistive load inverter case, the curves seems to be laterally shifted to higher temperatures due to the static power dissipation in this circuit.

With increasing rise and fall times at the drain, the current will flow for more time, which will result in more heating inside the device. It will also decrease the time available per cycle for the device to cool down, as the clock period is kept constant. This should explain the increase in lattice temperature with increasing rise and fall times at the drain. So if we keep
input rise and fall times, which are determined by the input load, to be a constant, the lattice temperature increases with the increasing output load.

![Graph showing temperature rise and fall times](image)

(a) Fitting for the peak values (adjusted signals)  (b) Fitting for the valley values (adjusted signals)

Figure 7.12 – Fitting of the variations with a polynomial function (peak) and linear function (valley)

### 7.2.3 Varying the ambient temperature

In this experiment we tried to vary the ambient temperature and studied how it affects the self heating of the device. The period was kept constant at 200ps, rise and fall times were 10ps at the gate and 25ps at the drain.

![Graph showing variation with ambient temperature](image)

(a) Variation of the peak and valley temperatures of the final oscillations with ambient temperature  (b) Variation of the temperature rise (ΔT_{max}) with the ambient temperature

Figure 7.13 – Variations with the ambient temperature

The lattice temperature increases with the ambient temperature. The dependence is linear with a slope slightly higher than 1 (fig. 7.13a). It makes more sense to study the variation of the temperature rise in the device than the actual temperature, as can be observed in fig. 7.13b. A fitting of the same was attempted, to better understand the trend. As can be observed from fig. 7.14, the maximum rise of lattice temperature follows a 2\textsuperscript{nd} order relation with the ambient temperature. The increase of the temperature rise with increasing ambient
temperature might be due to the fact that the lattice attains more energy at higher temperatures, which in turn will increase the probability of electron-lattice collisions.

![Figure 7.14 – Fitting of the variations of temperature rise with a polynomial functions](image)

(a) Fitting for the peak values  
(b) Fitting for the valley values

7.2.4 An Extreme Case

For the sake of illustration, an extreme case of operating the inverter was simulated. The period was chosen to be 200ps, rise and fall times at the gate was 10ps and at drain was 95ps. The gate and drain signals had a delay of 15ps between them. Even with the huge load, which makes the output rise and fall times approximately 95% of the half period, the temperature rise inside the device is less than 20K.

![Figure 7.15 – An extreme case, with huge output load](image)

(a) Initial variation of lattice temperature.  
(b) Variation of lattice temperature

7.2.5 Remarks

From the experiments conducted using the CMOS inverter, it was observed that the temperature rise due to the self heating is much lower during its operation. It was observed that the lattice temperature increased with increasing clock frequency of the applied signal and
with increasing output loads. Also the temperature rise increased with increasing ambient
temperature. In case of high frequency operation of the device, it is very important to ensure
sufficient thermal conduction from the device in a circuit, as the temperature rise in the de-
vice is more due to incomplete removal of the generated heat in the device in a clock cycle,
rather than due to increased heat generation per clock cycle.

7.3 Gate contact/Source(Drain) Contact Spacing Variation

In the previous chapter, when we were varying the contact spacing, we mentioned that the
increase of parasitic capacitance is one of the concerns when bringing the contacts closer.
At that point this was ignored, but now we will try to explore more about this using some
transient hydrodynamic simulations.

A setup as shown in the inset of fig. 7.16 was used to estimate the propagation delay of the
device. The switch to the dc supply was opened immediately after applying the gate pulse,
allowing the drain voltage to float freely. As can be observed, right after the application of
the gate pulse, the device with the longest spacing responds faster. But things start to change
later on. If one defines the propagation delay to be the time between the 50% of input to
50% of output, we find that trend is reversed. The device with the closest contacts seems to
be faster. This opposite trend can be explained if we refer back to variation of the saturated
drain current in the device (fig. 6.19b) which is decreasing with increasing spacing. The delay as measured above is a function of both parasitic capacitance and the drain current. To verify, we tried to estimate the parasitic capacitance with the measured propagation delay and drain current using the expression,

\[ Q = I \times T_{pd} = C_{par} \times V_d \]  \hspace{1cm} (7.2)

where \( Q \) is the charge stored, \( T_{pd} \) is the measured propagation delay, \( I \) is the saturated drain current, \( C_{par} \) is the parasitic capacitance and \( V_d \) is the dc drain voltage. This is a very vague approximation, as here we assumed a constant dc drain voltage and that the same saturated drain current will flow during the entire time, both of which are not completely true. But this will be sufficient to get an idea of the parasitic capacitance and to understand the trend. The calculated capacitance is plotted in fig. 7.16c. As can be observed, the parasitic capacitance is actually decreasing with the increasing contact spacing as predicted earlier[6]. To confirm the observed trends we tried to perform simulations using a CMOS inverter, in which the contact spacing was varied in both PMOS and NMOS. The observed inverter propagation delay is plotted in fig. 7.16d. Similar trends were also observed for the inverter propagation delays.

7.4 Effect of Surroundings in a Circuit

We investigated the effect of the surroundings of the device in a circuit. In modern circuits, it is often possible to find many devices close to each other when they are interconnected, or in some cases isolated by trench isolations. To consider both cases, we made use of two different structures, as shown in fig. 7.17. For the case in which the device is supposed to be surrounded by similar devices, we simply extended the SOI to either sides. Both quasistationary and transient simulations were performed.

Neither the lattice temperature nor the heat generation is much affected by the surroundings. But in the transient simulation, it was observed that the cooling of the device is different for each structure. The device without its surroundings (legend: device only) showed much faster cooling compared to other cases. All the three structures showed similar initial cooling, but once past ~50% point, the effect of the surroundings seems to play a bigger role.
The trench isolated structure showed the slowest cooling down, which can be attributed to the lower thermal conductivity of silicon dioxide. But this should not manifest itself in the high frequency operation of the devices, because as long as the period of the input is small enough, all three structures behave similarly (fig. 7.18d with input frequency of 5GHz). But for an operation at lower frequency, with period \( \geq 2\text{ns} \), this should cause much higher net temperatures in the device.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure7.18}
\caption{Effect of surroundings.}
\end{figure}

\subsection{7.5 Observations}

From the transient simulations performed, we were able to study the temporal behaviour of self heating of the device. This provides much valuable information when designing high frequency circuits, and helps the designer to choose the devices depending on the frequency it is supposed to operate. The fact that the temperature in the device is not constant during its operation can have great implications. More detailed studies might be required to understand whether such a varying temperature somehow affects the operation and functionality in a circuit. Also it was found even though making the gate and source/drain contacts closer increases the parasitic capacitances, the propagation delay seems to be improved due to effect of lower currents at higher lattice temperatures.
Chapter 8

Conclusion and Future Work

In the course of the four months on this project, we were able to tweak the Sentaurus TCAD hydrodynamic simulation parameters to obtain results comparable to the Monte Carlo reported values. We were able to perform a detailed analysis of the self heating in bulk and FDSOI MOSFETs, using both steady state and transient simulations.

We illustrated the importance of having true thermal boundary conditions to get meaningful results from the simulations. We also explored the different structural modifications which make the device more thermally efficient. We found that the temperature rise in the device due to self heating increases in shorter devices with thicker BOX. But it was also found that even though longer devices reported lower temperatures, they took more time to cool down to ambient temperature upon the removal of the input signal. We also demonstrated that placing the gate and source/drain contacts closer to each other is beneficial both from a thermal and electrical perspective. Contrary to prior arguments, we found that the propagation delay of the device improves with decreasing spacing. We found that even though the parasitic capacitance increases with decreasing spacing, the improvement in the drain current due to reduced self heating helps decrease the propagation delay.

Another important contribution of this work is the exploration of the temporal behaviour of self heating, which was performed for the first time, to the best of our knowledge. We found that the temperature inside the device follows the input signal very closely and varied quite a lot during the operation of the device. Also the drain current was found to follow this variation of the temperature. This could have various implications in a circuit, where the varying temperatures might induce false behaviours during the transient operation of the circuit. We explored the dependence of the self heating with the type of the signal and circuit parameters like the load being driven. It was found that the temperature mainly depends on the input and output loads of the device, and will increase with increasing output to input load ratio. Also it was found that it is very essential to provide sufficient cooling to the device, especially when it is operating at very high frequencies, as in this case the net temperatures are determined mostly by the insufficient cooling of the device.
At this point, it is quite important to ponder about the accuracy of hydrodynamic simulations. It is known that hydrodynamic simulations overestimates the drain current when compared to measured values\cite{45}. Also this model fails to take into account the phonon dispersion spectra of the temperature rise inside the device. Based on theoretical formulations \cite{7,8}, the interconversion between the acoustic and optical phonons inside the device and the mismatch between the generation rates of the phonons and the decay rates/group velocities of the generated phonons is supposed to play a significant role in the creation of the hot spots inside the device. None of these were considered in our simulations. As a further extension of our work, we need to validate our findings with an electro-thermal Monte Carlo simulator.

Further extensions can be included to better understand this phenomenon. Some of the worthwhile extensions which could be attempted are to perform process simulations to include advanced doping technologies like halo, graded substrate doping etc. These will help bring the simulation results more closer to the real device. Also performing some interconnect simulations will help understand the thermal dissipations in the interconnects and help improve the understanding from the circuit level.
Appendix A

Physics used in Sentaurus TCAD

The simulations were performed with Sentaurus device tool. A brief review of the physics, relevant to our simulations, will be described in this chapter. For more details and the values of the default parameters please refer to the user guide of Sentaurus device[58].

A.1 Electrode Definitions

The electrical boundary conditions are very important for obtaining true results from the device simulation. In Sentaurus device simulation tool, it is possible to define different electrical boundary conditions for each boundary, which has been later exploited to make the simulation complexity tractable. In our simulations we mainly used contact on insulator for the gate contact and ohmic contacts for the rest of the contacts, which were later replaced with resistive contacts to make the simulation more compact.

A.1.1 Ohmic Contacts

The default electrical boundary condition is that of an Ohmic contact which offers 0.001 $\Omega$ contact resistance when connected to an external circuit and zero elsewhere. Charge neutrality and equilibrium are assumed at Ohmic contacts,

$$n_0 - p_0 = N_D - N_A \quad (A.1)$$

$$n_0 p_0 = n_{i, eff}^2 \quad (A.2)$$

The carrier densities can be expressed as the following for Boltzmann statistics:

$$\phi = \phi_F + \frac{kT}{q} \text{asinh} \left( \frac{N_D - N_A}{2n_{i, eff}} \right) \quad (A.3)$$

$$n_0 = \sqrt{\frac{(N_D - N_A)^2}{4} + n_{i, eff}^2} + \frac{N_D - N_A}{2}, \quad p_0 = \sqrt{\frac{(N_D - N_A)^2}{4} + n_{i, eff}^2} - \frac{N_D - N_A}{2} \quad (A.4)$$
where \( n_0, p_0 \) are the electron and hole equilibrium concentrations and \( \Phi_F \) is the Fermi potential at the contact, which is equal to the applied voltage in case the contact resistance is zero. The tool uses the following boundary conditions:

\[
J_n \cdot \hat{n} = q v_n (n - n_0), \quad J_p \cdot \hat{n} = -q v_p (p - p_0)
\] (A.5)

where \( v_n \) and \( v_p \) are the electron and hole recombination velocities. [58]

A.1.2 Contacts on Insulators

This type of contact is used for the gate contact of the devices. In this case the electrostatic potential is given by,

\[
\phi = \phi_F - \Phi_{MS}
\] (A.6)

where \( \Phi_{MS} \) is the workfunction difference between the metal and the intrinsic reference semiconductor, which can be specified in the command file for Sentaurus device. [58]

A.1.3 Resistive Contacts

The default value of zero contact resistance for an ohmic contact can be changed in Sentaurus device. The tool provides two different ways of performing this: by defining a lumped resistance at the contact (unit: \( \Omega \)) or by defining a distributed resistance (unit: \( \Omega \text{cm}^2 \)). We used the second approach for the source and drain contacts.

In this case of distributed resistance, the boundary conditions get modified as,

\[
\hat{n} \cdot (J_p(\phi_F) + J_D(\phi_F)) d s = \frac{(V_{\text{applied}} - \phi_F)}{R_d}
\] (A.7)

where \( s \) is the contact area used in the simulations to compute the total current through the contact [58].

A.2 Thermal Boundary Conditions

Defining the thermal boundary conditions is important for solving the thermal equations when using a hydrodynamic simulation approach. The different thermal boundary conditions possible in Sentaurus device are,

- Thermally insulating surfaces.

\[
\kappa \hat{n} \cdot \nabla T = 0
\] (A.8)

This kind of boundary condition is assumed usually at the gate of the device, owing to the factor that in a real circuit, the gate is going to be buried under many layers of metal and oxide, which makes any cooling at the gate almost impossible. Also at the lateral sides of the 2D structure of the device, an insulating boundary condition is usually assumed to account for the factor that the device might be situated between many other devices, which could be having similar thermal behaviour in the circuit.
• Nonhomogeneous Neumann boundary conditions

\[ \kappa_n \cdot \nabla T = \frac{T_{\text{ext}} - T}{R_{th}} \quad (A.9) \]

where \( R_{th} \) is the external thermal resistance, which can be specified in the command file.

• Dirichlet conditions

\[ T = T_{\text{ext}} \quad (A.10) \]

This is a special case of the earlier one in which \( R_{th} \to 0 \), which is also the default value of \( R_{th} \). This corresponds to a case of an ideal heat sink.

The value of the \( R_{th} \) can be specified either as surface resistance (unit: \( cm^2KW^{-1} \)) or surface conductance (unit: \( cm^{-2}K^{-1}W \)) which specifies \( 1/R_{th} \). The external temperature can also be defined in the command file. The declaration of the thermodes makes the lattice temperature and the lattice heat fluxes at the contacts to be stored in the output plot file [58].

### A.3 Carrier Transport

Sentaurus device simulation tool provides different carrier transport models. The continuity equations can be expressed as,

\[ \nabla \cdot J_n = q R_{net} + q \frac{\partial n}{\partial t}, \quad \nabla \cdot J_p = q R_{net} + q \frac{\partial p}{\partial t} \quad (A.11) \]

where \( R_{net} \) is the net recombination rate, \( J_n \) and \( J_p \) is the electron and hole current densities respectively and \( n \) and \( p \) are the electron and hole densities respectively. The transport models differ in how they model the current densities. The different models available are:

- Drift-diffusion
- Thermodynamic
- Hydrodynamic
- Monte Carlo

We mostly use hydrodynamic approaches in our simulations, but this report will also look into the drift-diffusion model also, for completeness.

#### A.3.1 Drift-Diffusion

This is effectively an isothermal model, which describes the current densities with the components from the drift of the carriers due to the electric fields and the diffusion of the carriers due to concentration gradients.

\[ J_n = \mu_n (n \nabla E_C - 1.5nkT \nabla \ln m_n) + D_n (\nabla n - n \nabla \ln \gamma_n) \quad (A.12) \]
\[ J_p = \mu_p (p \nabla E_V - 1.5 n k T \nabla \ln m_p) + D_p (\nabla p - p \nabla \ln \gamma_p) \]  
(A.13)

The relation between the diffusivities \( D_n \) and \( D_p \) and the mobilities are given by Einstein’s relation,

\[ D_{n/p} = k T \mu_{n/p} \]  
(A.14)

When eqn (A.14) holds, the eqns (A.12) and (A.13) can be simplified as,

\[ J_n = -n q \mu_n \nabla \Phi_n \]  
(A.15)

\[ J_p = -p q \mu_p \nabla \Phi_p \]  
(A.16)

where \( \Phi_n \) and \( \Phi_p \) are the electron and hole quasi-Fermi potentials respectively [58].

### A.3.2 Hydrodynamic

This model accounts for the energy transport of the carriers. The current densities are given by,

\[ J_n = \mu_n (n \nabla E_C + k T_n \nabla n - n k T_n \nabla \ln \gamma_n + \lambda_n f_n^{id} k n \nabla T_n - 1.5 n k T_n \nabla \ln m_n) \]  
(A.17)

\[ J_p = \mu_p (p \nabla E_V + k T_p \nabla p - p k T_p \nabla \ln \gamma_p + \lambda_p f_p^{id} k p \nabla T_p - 1.5 p k T_p \nabla \ln m_p) \]  
(A.18)

The first term models the contribution due to the variation of electrostatic potentials, electron affinity and band gap. The remaining terms models the effect due to the gradient of concentration, the carrier temperature gradients and the spatial variations of effective masses. For Boltzmann statistics, which is the default one in Sentaurus device, \( \gamma_n = \gamma_p = \lambda_n = \lambda_p = 1 \) [58].

### A.4 Temperature Equations

It is possible to calculate three different temperatures in Sentaurus device: lattice temperature, electron and hole temperatures. The different approaches available in the tool to calculate the same are as below [58]:

- Lattice temperature computed from the total dissipated heat assuming the temperature to be constant in the device.
- Lattice temperature computed non-uniformly using the default model or the thermodynamic or the hydrodynamic models.
- The carrier temperatures computed using the hydrodynamic model

In our simulations we used the approach the lattice and the carrier temperatures to be calculated using the hydrodynamic model.
A.4.1 Hydrodynamic model

Sentaurus device uses a simpler formulation which follows the work of Bløtekjær[44] and Stratton[43], but without any convective terms. In addition to the Poisson’s equation and continuity equation, the tool solves up to three additional energy balance equations for the lattice temperature and carrier temperatures.

\[
\frac{\partial W_n}{\partial t} + \nabla \cdot S_n = J_n \cdot \nabla E_C + \frac{dW_n}{dt} \bigg|_{coll} \\
\frac{\partial W_p}{\partial t} + \nabla \cdot S_p = J_p \cdot \nabla E_V + \frac{dW_p}{dt} \bigg|_{coll} \\
\frac{\partial W_L}{\partial t} + \nabla \cdot S_L = \frac{dW_L}{dt} \bigg|_{coll}
\]

where the energy fluxes are modelled as,

\[
S_n = -\frac{5r_n\lambda_n}{2} \left( \frac{kT_n}{q} J_n + f_n^{hf} \kappa_n \nabla T_n \right) \\
S_p = -\frac{5r_p\lambda_p}{2} \left( \frac{kT_p}{q} J_p + f_p^{hf} \kappa_p \nabla T_p \right) \\
S_L = -\kappa_L \nabla T_L
\]

\[
\kappa_n = \frac{k^2}{q} n\mu_n T_n \\
\kappa_p = \frac{k^2}{q} p\mu_p T_p
\]

where the default values of \( r_n \) and \( r_p \) is 0.6 and that of \( f_n^{hf} \) and \( f_p^{hf} \) is 1.

The collision terms are expressed as,

\[
\frac{\partial W_n}{\partial t} \bigg|_{coll} = -H_n - \xi_n \frac{W_n - W_{n0}}{\tau_{en}} \\
\frac{\partial W_p}{\partial t} \bigg|_{coll} = -H_p - \xi_p \frac{W_p - W_{p0}}{\tau_{ep}} \\
\frac{\partial W_L}{\partial t} \bigg|_{coll} = H_L + \xi_n \frac{W_n - W_{n0}}{\tau_{en}} + \xi_p \frac{W_p - W_{p0}}{\tau_{ep}}
\]

where \( H_n, H_p \) and \( H_L \) are the energy gain/loss due to generation/recombination,

\[
H_n = 1.5kT_n(R_{SRH}^{net} + R_{rad}^{net} + R_{trap}^{net}) - E_{g,eff}(R_n^A - G_n^{ij}) - \alpha(\hbar\omega - E_{g,eff})G^{opt} \\
H_p = 1.5kT_p(R_{SRH}^{net} + R_{rad}^{net} + R_{trap}^{net}) - E_{g,eff}(R_p^A - G_p^{ij}) - (1 - \alpha)(\hbar\omega - E_{g,eff})G^{opt}
\]
\[ H_L = [R_{SRH}^{net} + 0.5(R_{trap}^{tr, net} + R_{trap}^{tr, net})](E_{R, eff} + 1.5kT_n + 1.5kT_p) \] (A.32)

where \( R_{SRH}^{net} \) is the Shockley-Read-Hall recombination rate, \( R_{rad}^{net} \) is the radiative recombination rate, \( R_n^A \) and \( R_p^A \) are Auger recombination rates, \( \hbar \omega \) is the photon energy, \( \alpha \) is a dimensionless parameter which describes the surplus energy of the photon splits between the bands, \( G_{n}^{ii} \) and \( G_{p}^{ii} \) are the impact ionization generation rates, \( R_{trap}^{tr, net} \) and \( R_{trap}^{tr, net} \) are the recombination rates through the trap levels and \( G_{opt}^{opt} \) is the optical generation rate.

The energy densities are given by,

\[ W_n = nw_n = n \left( \frac{3kT_n}{2} \right) \] (A.33)

\[ W_p = pw_p = p \left( \frac{3kT_p}{2} \right) \] (A.34)

\[ W_L = c_L T \] (A.35)

The parameters \( \xi \) improve the numerical stability and is given by,

\[ \xi_n = 1 + \frac{n_{min}}{n} \left( \frac{n_0}{n_{min}} \right)^{\max[0,(T-T_n)/100K]} \] (A.36)

where the parameters \( n_{min} \) and \( n_0 \) are adjustable. The equation for the holes is similar [58].

A.5 Intrinsic Density and Bandgap Narrowing

The intrinsic density for undoped semiconductors is given by,

\[ n_i(T) = \sqrt{N_C(T)N_V(T)exp \left( -\frac{E_g(T)}{2kT} \right)} \] (A.37)

and the effective intrinsic density with doping dependent bandgap narrowing is given by,

\[ n_{i, eff} = n_i exp \left( \frac{E_{bg n}}{2kT} \right) \] (A.38)

Sentaurus device provides the users a choice of different models for the bandgap narrowing:

- Bennet Wilson model
- delAlamo model
- Slotboom model
- JainRoulston model
- Based on a user specified table
For our simulations, we used the Slotboom model for bandgap narrowing.

The bandgap is modelled as below in Sentaurus. It includes a temperature dependence and doping dependent band gap narrowing.

\[ E_g(T) = E_g(0) - \frac{a T^2}{T + \beta} \]  
\[ (A.39) \]

where \( E_g(0) \) is the bandgap energy at 0K, and is given by,

\[ E_g(0) = E_{g,0} + \delta E_{g,0} \]  
\[ (A.40) \]

where each of the component of eq\[A.40\] are different in each model. The effective band gap is given by,

\[ E_{g,eff}(T) = E_g(T) - E_{bgn} \]  
\[ (A.41) \]

and the electron affinity is given by,

\[ \chi(T) = \chi_0 + \frac{(\alpha + \alpha_2)T^2}{2(T + \beta + \beta_2)} + Bgn2Chi \cdot E_{bgn} \]  
\[ (A.42) \]

where \( \chi_0 \) and \( Bgn2Chi \) are adjustable parameters.

The different models handle the bandgap narrowing differently,

\[ E_{bgn} = \Delta E_g^0 + \Delta E_g^{Fermi} \]  
\[ (A.43) \]

where the first term is determined by the model being used, and the second is an optional term to account for carrier statistics [58].

**A.5.1 Slotboom Model**

In this model, \( \Delta E_g^0 \) is modelled as [58],

\[ \Delta E_g^0 = E_{ref} \left[ \ln \left( \frac{N_{tot}}{N_{ref}} \right) + \sqrt{\left( \ln \left( \frac{N_{tot}}{N_{ref}} \right) \right)^2 + 0.5} \right] \]  
\[ (A.44) \]

**A.6 Mobility**

The unified mobility model proposed by Klaassen [46], was used for the simulations. This model combines the effects of temperature, electron-hole scattering, screening of ionized impurities by charge carriers and clustering of impurities. The different contributions are combined using Matthiessen’s rule. [58]

\[ \frac{1}{\mu_{i,b}} = \frac{1}{\mu_{i,L}} + \frac{1}{\mu_{i,D_Aeh}}, \]  
\[ (A.45) \]

where \( \mu_{i,b} \) is the total mobility, \( \mu_{i,L} \) is the contribution from the lattice scattering and \( \mu_{i,D_Aeh} \) is the contribution from the other scatterings.
The lattice scattering contribution is modelled as,
\[ \mu_{i,L} = \mu_{i,max} \left( \frac{T}{300K} \right)^{-\theta_i}, \quad (A.46) \]
where \( \theta_i \) has a default value of 2.285.

The second contribution is given by
\[ \mu_{i,D\DeltaHeh} = \mu_{i,N} \left( \frac{N_{i,sc}}{N_{i,sc,eff}} \right)^{\alpha} + \mu_{i,c} \left( \frac{n + p}{N_{i,sc,eff}} \right) \quad (A.47) \]
with:
\[ \mu_{i,N} = \frac{\mu_{i,\text{max}}^{2} T}{\mu_{i,\text{max}} - \mu_{i,\text{min}}} \left( \frac{T}{300K} \right)^{3\alpha_i - 1.5} \quad (A.48) \]
\[ \mu_{i,c} = \frac{\mu_{i,\text{max}} \mu_{i,\text{min}}^{2}}{\mu_{i,\text{max}} - \mu_{i,\text{min}}} \left( \frac{300K}{T} \right)^{0.5} \quad (A.49) \]
for the electrons:
\[ N_{e,sc} = N_{D}^{*} + N_{A}^{*} + p \quad (A.50) \]
\[ N_{e,sc,eff} = N_{D}^{*} + G(P_e)N_{A}^{*} + f_e \frac{p}{F(P_e)} \quad (A.51) \]
and for holes:
\[ N_{h,sc} = N_{A}^{*} + N_{D}^{*} + n \quad (A.52) \]
\[ N_{h,sc,eff} = N_{A}^{*} + G(P_h)N_{D}^{*} + f_h \frac{n}{F(P_h)} \quad (A.53) \]
The effects of clustering of donor and acceptors at ultrahigh concentrations are described as,
\[ N_{D}^{*} = N_{D,0} Z_{D} = N_{D,0} \left[ 1 + \frac{N_{D,0}^{2}}{c_D N_{D,0}^{2} + N_{D,ref}^{2}} \right] \quad (A.54) \]
\[ N_{A}^{*} = N_{A,0} Z_{A} = N_{A,0} \left[ 1 + \frac{N_{A,0}^{2}}{c_A N_{A,0}^{2} + N_{A,ref}^{2}} \right] \quad (A.55) \]
The screening factor functions \( G(P_i) \) and \( F(P_i) \) are given as,
\[ F(P_i) = \frac{0.7643 P_i^{0.6478} + 2.2999 + 6.5502 (m_i^* / m_j^*)}{P_i^{0.6478} + 2.3670 - 0.8552 (m_i^* / m_j^*)} \quad (A.56) \]
and
\[ G(P_i) = 1 - \frac{0.89233}{0.41372 + P_i \left( m_j/m_i \right)^{0.28227}} \left[ 0.19778 + P_i \left( m_j/m_i \right)^{0.72169} \right]^{1.80618} \quad (A.57) \]
The screening parameter, \( P_i \) is given by,
\[ P_i = \left[ \frac{f_{CW}}{3.97 \times 10^{13} \text{cm}^{-2} N_{i,sc}^{2/3}} + \frac{f_{BH}}{1.36 \times 10^{20} \text{cm}^{-3} m_j^*} \right] \left( \frac{T}{300K} \right)^{2} \quad (A.58) \]
A.6.1 High field saturation

The actual mobility is modified with the Caughey-Thomas expression in the case of high field.

\[
\mu(F) = \frac{(\alpha + 1)\mu_{\text{low}}}{\alpha + \left[1 + \left(\frac{(\alpha + 1)\mu_{\text{low}}F_{\text{hfs}}}{v_{\text{sat}}}\right)^\beta\right]^{1/\beta}}
\]  

(A.59)

where \(\mu_{\text{low}}\) is the low field mobility as expressed in previous section and

\[
\beta = \beta_0\left(\frac{T}{300K}\right)^{\beta_{\text{exp}}}
\]

(A.60)

where \(\beta_{\text{exp}}\) = 0.66 (for electrons) and 0.17 (for holes) and \(\beta_0\) = 1.109 (for electrons) and 1.213 (for holes) [58].

The driving force, \(F_{\text{hfs}}\) calculated using the CarrierTempDrive model, which is the default one for hydrodynamic simulations,

\[
F_{\text{hfs},n} = \sqrt{\max(w_n - w_0, 0)} \frac{\tau_{e,n} q \mu_n}{\tau_n}
\]

(A.61)

where \(w_n = 3kT_n/2\) is the average electron thermal energy, \(w_0 = 3kT/2\) is the equilibrium thermal energy and \(\tau_{e,n}\) is the energy relaxation time [58].

In our simulations we also included the effects of interface scattering on the mobility of the carriers in the material. This is modelled using the Lombardi model, details of which can be found in the Sentaurus device user guide [58].

A.7 Recombination

Only Shockley-Read-Hall recombination was considered in our simulations. Both Auger and band-to-band recombination was ignored.

A.7.1 Shockley-Read-Hall(SRH) Recombination

The SRH recombination is modelled as,

\[
R_{\text{net}}^{\text{SRH}} = \frac{np - n_{i,\text{eff}}^2}{\tau_p (n + n_i) + \tau_n (p + p_i)}
\]

(A.62)

with:

\[
n_i = n_{i,\text{eff}} \exp\left(\frac{E_{\text{trap}}}{kT}\right)
\]

(A.63)

and

\[
p_i = n_{i,\text{eff}} \exp\left(-\frac{E_{\text{trap}}}{kT}\right)
\]

(A.64)
where $E_{\text{trap}}$ is the difference between the defect level and intrinsic level. The silicon default value for $E_{\text{trap}} = 0$.

The lifetimes, $\tau_p$ and $\tau_n$ are modeled as a product of a doping-dependent, field-dependent, and temperature-dependent factor:

$$\tau_c = \tau_{dop} \frac{f(T)}{1 + g_c(F)}$$  \hspace{1cm} (A.65)

The doping dependence of the lifetimes in Sentaurus is modeled as,

$$\tau_{dop}(N_{A,0} + N_{D,0}) = \tau_{min} + \frac{\tau_{max} - \tau_{min}}{1 + \left(\frac{N_{A,0} + N_{D,0}}{N_{ref}}\right)^\gamma}$$  \hspace{1cm} (A.66)

The temperature dependence is modeled as,

$$\tau_{SRH}(T) = \tau_{SRH}(300K) \left(\frac{T}{300K}\right)^{T_a}$$  \hspace{1cm} (A.67)

with $T_a = 3/2$ for Silicon (default) [58].
Appendix B

Distribution of Lattice Temperature and Heat Generation in 25nm FDSOI

Figure B.1 – Heat generation inside the device for different gate voltages and drain voltage=1.2V
Figure B.2 – Temperature distribution inside the device for different gate voltages and drain voltage=1.2V
Appendix C

Variation of Heat generation and Temperature Rise in Terms of Applied Voltages in 25nm FDSOI

C.1 Variation with drain voltage

(a) Variation of max heat generation

(b) Variation of max heat generation/drain current

Figure C.1 – Variation with drain voltage for different BOX thicknesses

(a) Variation of max lattice temperature rise

(b) Variation of max temperature rise/drain current

Figure C.2 – Variation with drain voltage for different BOX thicknesses
C.2 Variation with gate voltage

(a) Variation of max heat generation
(b) Variation of max heat generation/drain current

Figure C.3 – Variation with gate voltage for different BOX thicknesses

(a) Variation of max lattice temperature rise
(b) Variation of max temperature rise/current

Figure C.4 – Variation with gate voltage for different BOX thicknesses
Appendix D

Temperature Distribution in 25nm FDSOI as a Function of Time

Figure D.1 – Temperature distribution inside the device during the time in which the pulse is applied at the gate. The time scale used is the same as in fig 7.1.
Figure D.2 – Temperature distribution inside the device during the time in which the pulse is not applied at the gate. The time scale used is the same as in Fig. 7.1.
Bibliography


