9-9-2009

Simulation of INSB Devices using Drift-Diffusion Equations

Edin Sijercic
Portland State University

DISSERTATION APPROVAL

The abstract and dissertation of Edin Sijercic for the Doctor of Philosophy in Electrical and Computer Engineering were presented September 9, 2009, and accepted by the dissertation committee and the doctoral program.

COMMITTEE APPROVALS:

Branimir Pejcinovic, Chair

Robert Daasch

James Morris

Christof Teuscher

Rolf Koenenkamp
Representative of the Office of Graduate Studies

DOCTORAL PROGRAM APPROVAL:

Malgorzata Chrzanowska-Jeske, Chair
Department of Electrical and Computer Engineering

ABSTRACT

An abstract of the dissertation of Edin Sijercic for the Doctor of Philosophy in Electrical and Computer Engineering presented September 9, 2009.

Title: Simulation of InSb Devices Using Drift-Diffusion Equations

Silicon technology has for several decades followed Moore's law. Reduction of feature dimensions has resulted in constant increase in device density which has enabled increased functionality. Simultaneously, performance, such as circuit speed, has been improving. Recently, this trend is in jeopardy due to, for example, unsustainable increase in the processor power dissipation. In order to continue development trends, as outlined in ITRS roadmap, new approaches seem to be required once feature size reaches 10 – 20 nm range.

This research focuses on using III-V compounds, specifically indium-antimonide (InSb), to supplement silicon CMOS technology. Due to its low bandgap and high mobility, InSb shows promise as a material for extremely high frequency active devices operating at very low voltages. In this research electrical properties of InSb material are characterized and modeled with special emphasis on recombination-generation mechanisms. Device simulators based on drift-diffusion approach - DESSIS and nanoMOS – are

modified for InSb MOSFET design and analysis. To assess the quality of InSb MOSFET designs several figures of merit are utilized: I_{on}/I_{off} ratio, I-V characteristics, threshold voltage, drain induced barrier lowering (DIBL) and unity current gain frequency for different configurations and gate lengths. It is shown that significant performance improvement can be achieved in InSb MOSFETs through proper scaling. For example, extrapolated cutoff frequencies reach into THz range. Semi-empirical scaling rules that remedy short channel effects are proposed. Finally, quantum mechanical (QM) effects in InSb MOSFET and their effect on device performance are examined using nanoMOS device simulation program. It is found that nonparabolicity has to be properly modeled and that QM effects have a large effect on threshold voltage and transconductance and should be included when analyzing and designing deca-nanometer size InSb MOSFETs.

SIMULATION OF INSB DEVICES

USING DRIFT-DIFFUSION EQUATIONS

by

EDIN SIJERCIC

A dissertation submitted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY
in
ELECTRICAL AND COMPUTER ENGINEERING

Portland State University
2009

Acknowledgment

I would like to express gratitude to my advisor, Dr. Branimir Pejcinovic, who provided me with constant guidance and valuable advice, Intel Corporation, for their financial support of this project, Dr. Robert Chau and my parents.

Table of contents

Acknowledgment..... i

List of Tables v

List of Figures..... vi

1. Introduction and motivation..... 1

1.1 Problem Statement 1

1.2 Current Solutions 3

1.3. Potential Future Solutions 4

2. Research goals and contributions..... ...11

2.1 Research goals 11

2.2 Contributions 12

3. Material properties.....16

3.1 Band structure and density of states 17

3.2 2D electron concentration 22

3.3 Degeneracy and intrinsic concentration 25

3.4 Mobility and velocity saturation 28

3.5 Auger Recombination and Impact Ionization 30

3.6 Shockley-Reed-Hall (SRH) recombination 36

3.7 Band-to-Band tunneling 37

4. DESSIS Implementation..... 38

5. InSb diodes and Exclusion/extraction Mechanism..... 43

5.1 Extraction Diode 44

5.2 Exclusion/extraction Diode 46

6. Diode simulations and results..... 50

6.1 I-V characteristics 50

6.2 Differential resistance 51

7. Exclusion/extraction MOSFET..... 56

7.1 Device description 56

7.2. Short Channel effects 58

7.3 Simple Scaling results 60

7.4 Device scaling results 62

8. Quantum Mechanical effects..... 68

8.1 Non-parabolic quantum mechanical model 68

8.2 Impact of non-parabolicity on energy levels in

finite quantum wells..... 71

8.3 Dual gate InSb MOSFET device 74

8.4. Device Results 76

9. Conclusions..... 85

References.....89

List of Tables

Table 1: Electrical properties of common semiconductors 6

Table 2: Pros and Cons of the material properties of

InSb for the transistor applications..... 9

Table 3: Intrinsic concentration for different materials

at 77K and 300K..... 27

Table 4: Arora mobility model parameters..... 40

Table 5: Lombardi model parameters for InSb..... 41

Table 6: Threshold voltage [V_T] vs. V_{ds} and L_g 62

Table 7: Channel doping vs. L_g 63

List of Figures

Figure 1: Trends in the evolution of the high speed devices show exponential increase in the unity gain frequency which can lead to the prohibitive increase in the power density [1] Copyright by World Scientific Publishing Company... 1

Figure 2: Processor power trends show much faster increase of the leakage power over active power which may impede future generations of processors [3]. Copyright by Intel Corporation..... 3

Figure 3: Energy-delay product in Si MOSFETs, III-V QWFETs (or HEMTs) and carbon nanotube FETs; superior performance of the III-V devices [5]. Copyright by IEEE..... 7

Figure 4: InSb band structure [20]. Copyright by World Scientific..... 18

Figure 5: Electron concentration vs. Fermi level for three different approximations for density of states and Fermi integrals (parabolic, eq. (13) truncated at Fermi integral of 3/2 order and full eq. (13)) show significance of nonparabolicity..... 21

Figure 6: Comparison of different models for 2D electron concentration in InSb..... 25

Figure 7: np product in equilibrium vs. Fermi level shows a large variation of two decades.....28

Figure 8: Impact ionization; electron-hole pair is generated contributing to the leakage current in the reverse bias..... 31

Figure 9: Auger coefficient C_n for electrons in InSb as a function of electron concentration. It is decreasing with increasing concentration due to effective enlargement of the bandgap..... 33

Figure 10: Equilibrium Auger recombination and electron lifetime vs Fermi level: solid lines are from full band calculation, dashed from Beattie model in eq. 22 and 24. It reaches a maximum slightly above the bottom of the conduction band. Copyright by the American Institute of Physics..... 35

Figure 11: Mobility vs. vertical field plot shows rapid decrease in electron mobility with the vertical field..... 42

Figure 12: Extraction (top) and exclusion/extraction (bottom) diode cross section..... 44

Figure 13: Electron and hole concentrations for extraction and extraction/exclusion diodes under reverse bias V=-0.35 V..... 45

Figure 14: Band diagrams for $n^+ - \pi - \underline{p}^+ - p^+$ exclusion/extraction diode in equilibrium..... 47

Figure 15: Auger generation in n^+–p–p^+ and $n^+ - \pi - \underline{p}^+ - p^+$ diodes under reverse bias. Exclusion/extraction diode has reduced generation by more than a decade in p^+ region close to p-n junction..... 49

Figure 16: Leakage currents in $n^+ - \pi - \underline{p}^+ - p^+$ (exclusion/extraction) and n^+-π-p^+ (extraction) diodes show improvements in the reverse bias by addition of thin InAlSb layer..... 51

Figure 17: Differential Resistance: simulation vs. measured data for extraction diode (labeled "InSb") and exclusion/extraction diode (labeled" InAlSb"). Extraction diode shows relatively good fit with the measured data..... 53

Figure 18: Idealized cross section of the InSb MOSFET based on [6]..... 57

Figure 19: InSb MOSFET cross section, as used in DESSIS all units in μm..... 57

Figure 20: I_{ds} vs V_{gs} for different gate lengths; short channel effects become dominant at 300nm device..... 61

Figure 21: I_{on}/I_{off} vs. gate length show more than 3 decades in I_{on}/I_{off} separation, which is satisfactory performance; I_{off} is I_{ds} at zero volts, I_{on} at 1V..... 64

Figure 22: f_T vs. gate length and gate voltage V_{gs}. inSb shows very high values for the unity gain frequency..... 66

Figure 23: Extrapolated f_T demonstrates InSb is the most promising material for high speed applications. Numbers for Si and InP from [5]..... 67

Figure 24: 1eV rectangular quantum well ground states; Below 15nm nonparabolicity has to be taken into account..... 72

Figure 25: Comparison of parabolic and nonparabolic E-k diagram. Assumption of parabolicity results in large errors in energy..... 74

Figure 26: Dual gate InSb MOSFET cross section..... 75

Figure 27: I_{ds} vs. V_{gs} for V_{ds}=0.5V show large shift in the threshold voltage due to the nonparabolicity..... 77

Figure 28: Subthreshold I-V curves and subthreshold slope calculation; nonparabolicity slightly increases subthreshold slope..... 78

Figure 29: Effective electron mass along the channel (in units of free electron mass) is virtually constant, though different from the bottom of the Γ valley value..... 79

Figure 30: Variation of the potential in the vertical direction is within 10%..... 80

Figure 31: Max. Transconductance vs. body thickness shows substantial decrease in the transconductance with the reduced body thickness due to the mixing of the inversion layers..... 81

Figure 32: Electron concentration vs. vertical position for 5 and 15 nm thick device; maximum concentration occurs in the middle of the device due to the QM effects..... 82

Figure 33: Unity gain frequency is directly proportional to the body thickness due to the mixing of the inversion layers..... 84

1. Introduction and motivation

1.1 Problem statement

One of the most commonly used parameters to asses performance of the high-speed devices is the unity gain frequency. Trends in the recent years, for Silicon based MOSFETs, SiGe HBTs and InP HEMTs are shown in the Fig. 1 [1]

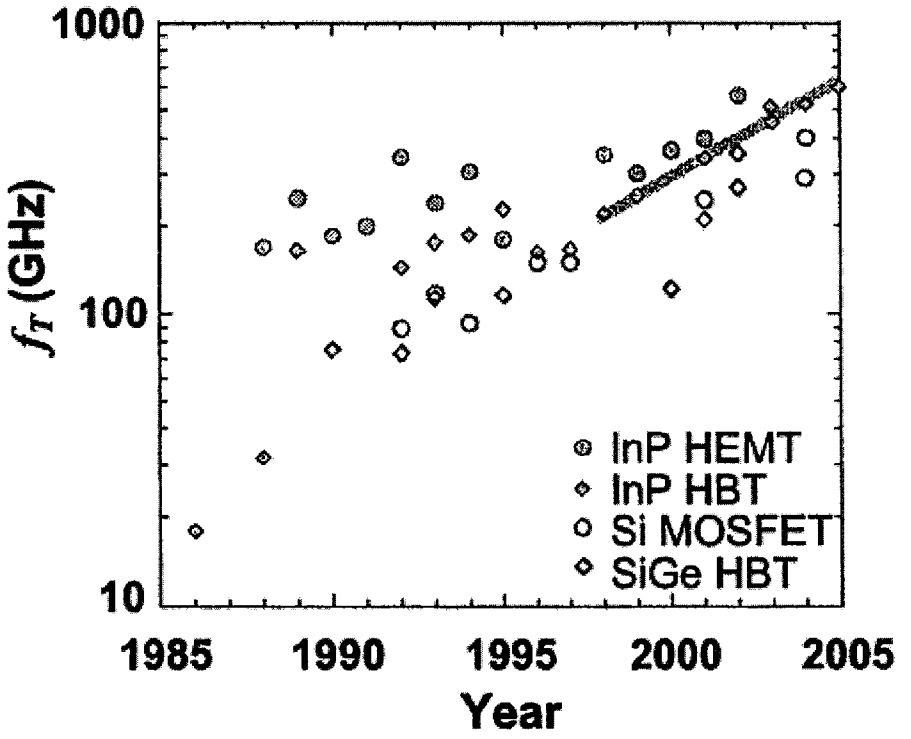

Figure 1: Trends in the evolution of the high speed devices show exponential increase in the unity gain frequency which can lead to the prohibitive increase in the power density [1] Copyright by World Scientific Publishing Company

In the past, increase in the unity gain frequency has led to the increase in the operation frequency. This increase, in turn, would lead to the increased power dissipation, for example through interconnect losses [2], increase in the leakage current and the nature of CMOS circuit operation. Also, reduction in the gate length, which is mainly responsible for the increase in the frequency shown in the graph, will enable larger device densities, leading to the larger power density. Currently, smallest commercial MOSFET gate length is 35 nm and HEMT 50 nm.

Increase in the processor power, active and leakage, is shown in the Fig. 2 [3]. Clearly, these figures illustrate that power is one of the main roadblocks to successfully maintaining Moore's law. In order to sustain product technology trends, as outlined in ITRS roadmap [4], new technologies will be required.

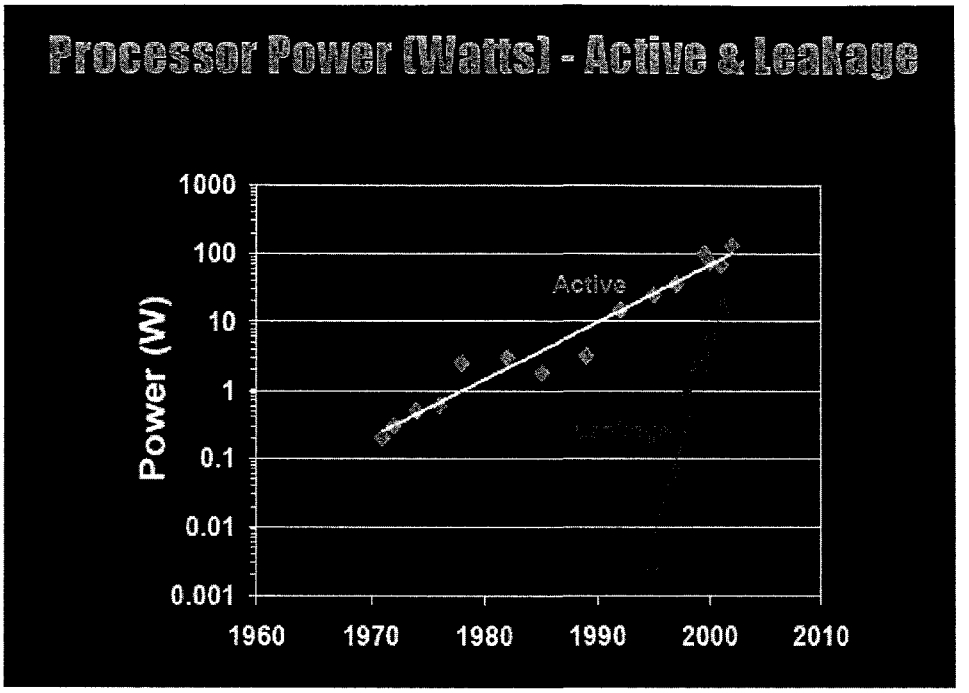

Figure 2: Processor power trends show much faster increase of the leakage power over active power which may impede future generations of processors [3]. Copyright by Intel Corporation

1.2. Current solution

Further increases in the frequency of the execution of the single core processor would lead to the enormous power dissipation, which becomes prohibitive. One way to avoid this "wall", which microprocessor industry has already adopted, is to increase throughput, i.e. parallelize execution by using multiple cores. Trade-off in this approach is diminishing return with increased number of cores. At some point in the future, adding cores would not increase performance substantially, given the Amdahl's "Law":

$$Max.\quad speedup = \frac{1}{(1-P)+\frac{P}{N}} \tag{1}$$

Where P is the fraction of the software code that can be parallelized and N is the number of cores. If number of cores tends to infinity, benefit or speed up will saturate at $\frac{1}{1-P}$. In real world applications, this saturation will occur fairly quickly. Also, this approach presents challenge to the software developers to massively parallelize code, increasing the value of P.

1.3 Potential future solutions

Alternative way would be to use materials other than silicon, which would enable higher current gain cut-off frequency f_T for the same gate length and lower power consumption. This approach is already subject of intense research, e.g. in [5, 6, 7, 8]. However, some of the drawbacks of these new materials include the lack of the suitable process infrastructure, and the cost of setting one up, since the semiconductor industry is currently very heavily invested in silicon manufacturing infrastructure. Manufacturing of the devices based on these new materials or device designs can be substantially more expensive than silicon. Other potential issues are increased leakage at the room temperature, requiring cooling or some other expensive technique to enable device operation. Also, some materials may have a very different performance for PMOS and NMOS devices, making CMOS logic useless, i.e.

slower device reducing overall performance or requiring larger area, rendering this new technology impractical. Hence, new logic technologies may be required, potentially increasing area and the cost.

For example, carbon nanotubes have been investigated for transistor applications [9, 10]. Although they do show some promising results, there are problems with mass production and control of the nanotube conductivity. Surface features, which are hard to control, can condition nanotube to be either a conductor or semiconductor rendering its use in practical circuits somewhat dubious.

Use of alternative materials is the subject of the research presented here. We have focused on III-V compounds, in particular indium antimonide, as an alternative to silicon. Among bulk III-V compounds, InSb shows the highest mobility ($7.8x10^4 cm^2V^{-1}s^{-1}$) and saturation velocity ($5x10^7$ cm/s) and has the smallest bandgap (0.17 eV). Hence, we expect InSb MOSFET devices to have much higher f_T than silicon based, for the same gate length. Also, due to its high electron mobility, "knee" voltage on the I-V curve could be attained at much lower drain voltage, compared to the silicon, allowing for a low power operation. Comparison of some important physical parameters of Si, GaAs, InAs and InSb is shown in the Table 1. In practical MOSFET devices it is inversion layer mobility that matters and it is usually significantly smaller than the bulk value, as discussed later. One exception are HEMT devices that can have channel mobility close to intrinsic material values [11]. Mobility can also

be manipulated by judicious use of strain as is done in modern Si MOSFETs [12, 13].

	Si	InSb	GaAs	InAs
Bandgap [eV]	1.1	0.175	1.43	0.354
effective electron mass	0.19	0.013	0.072	0.023
electron mobility [$cm^2V^{-1}s^{-1}$]	1,500	78,000	8,500	40,000
saturation velocity [$cm\ s^{-1}$]	10^7	$5*10^7$	10^7	$3.5*10^7$
Hole mobility	450	850	400	500

Table 1: Electrical properties of common semiconductors

In addition, Intel and QinetiQ have published work on quantum well transistors using indium antimonide [5]. This device shows promising trends as a substitute for silicon in the future MOSFETs. In particular, energy-delay product is substantially improved in InSb device over silicon based MOSFET, due to lower supply voltage and higher electron mobility, as shown in the Fig. 3 [5]. Energy-delay product is a useful metric for the efficiency of the devices, which represents energy required to run device at the particular speed. Lowering energy-delay product enables operation at the higher speed with the same energy, or alternatively, same speed operation at the lower energy [14].

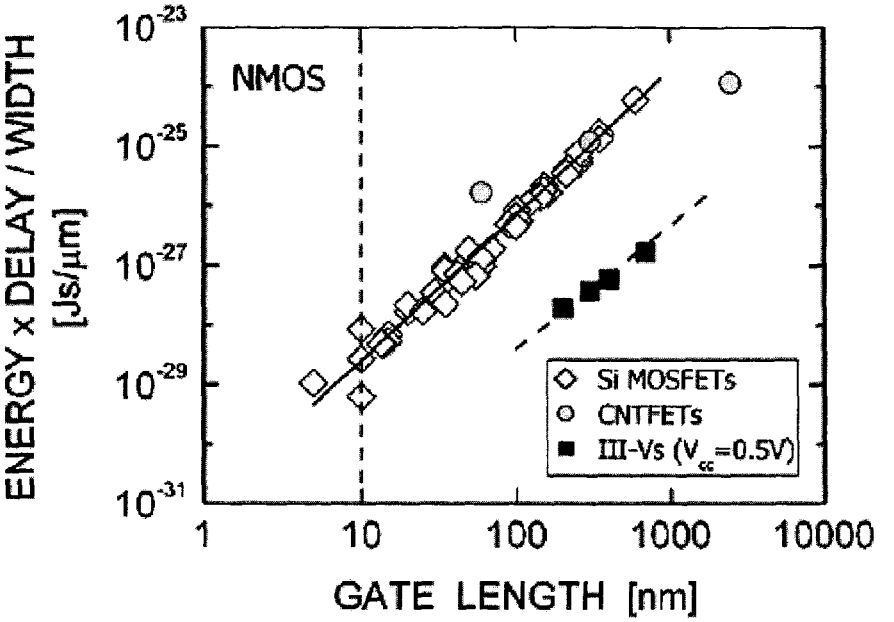

Figure 3: Energy-delay product in Si MOSFETs, III-V QWFETs (or HEMTs) and carbon nanotube FETs; superior performance of the III-V devices [5]. Copyright by IEEE

Exponential fits were generated for the graphs in the Fig. 3 and are given in eq. (2) – (4). In general, data points for the proposed solution should be below the silicon curve by at least factor of 2, in order to make it economically attractive.

If we choose 100 nm gate length, separation in energy-delay product between InSb and Si is more than a decade. This means that for the same gate length we can achieve better figure of merit, or alternatively, for the fixed energy delay product we can utilize larger gate length with InSb. Hence, Intel

researchers [5] conclude that InSb is the most promising material in this respect and that it could lead to significant reduction of dissipated power.

$$Si \quad MOSFET = 5.22 \cdot 10^{-32} \cdot e^{(\ln(L_g)2.586)} \tag{2}$$

$$CN \quad FET = 10^{-27} \cdot e^{(\ln(L_g))} \tag{3}$$

$$InSbHEMT = 1.871 \cdot 10^{-32} \cdot e^{(\ln(L_g)1.58)} \tag{4}$$

However, due to very small bandgap, InSb devices may have high leakage current at room temperature. For this reason, InSb devices were used primarily as photo-diodes at low temperatures (77K) but not as transistors for digital logic. This problem needs to be solved if InSb is to be considered as a supplement to silicon. Also, InSb hole mobility is about 100 times smaller than electron mobility at the room temperature and intrinsic material. This makes p-type device much slower than n-type, making CMOS technology with InSb inefficient, as explained earlier. New logic operation using n-type devices only should be employed and investigated. Alternatively, p-channel devices could be improved through strain engineering or by employing heterojunction structures. In this dissertation, however, our primary interest is in investigating n-channel devices.

Additional challenge in making the case for and investigating InSb transistor operation, which is the ultimate objective of this research and dissertation, is adequate modeling. High nonparabolicity and degeneracy

necessitate modification of the carrier density equations, as well as Schrödinger equation. Based on our literature survey, only limited effort has been spent in this direction. Detailed overview of the current state of the art is given at the beginning of each subsection.

Summary of the pros and cons of InSb discussed in our work is given in Table 2.

Pro	Con
High mobility	High electron intrinsic conc
High velocity sat	Large disparity between n and p mobility

Table 2: Pros and Cons of the material properties of InSb for the transistor applications

In the first part of the dissertation, we discuss research goals and metrics used to assess performance of the proposed device. In the second part we describe in detail important electrical properties of InSb which enable high performance MOSFET. In the third part DESSIS software implementation of this material is described. Since recombination-generation limits I_{on}/I_{off} ratio, in the fourth section we discuss these mechanisms in detail. Fifth part describes exclusion/extraction principle, which limits carrier generation in InSb device, enabling room temperature operation. Calibration and validation of this mechanism, through diode simulations, is described in the sixth section. Part seven discusses design of relatively large InSb MOSFETs and their performance and part eight deals with quantum mechanical effects in deca-

nanometer MOSFETs. Here we discuss necessary modification of the Schrödinger equation to account for the nonparabolicity and propose self-consistent methodology to calculate the effective mass and energy eigenvalues. This methodology is demonstrated on the dual-gate MOSFET device. Finally, we close with conclusions and make suggestions for future work.

2. Research goals and contributions

2.1 Research goals

The main research goals of this project are to develop appropriate material and device models and to examine feasibility and advantages of InSb MOSFET device over comparable silicon technology. To achieve that, following tasks have been completed:

1. characterize and model electrical properties of InSb,
2. demonstrate applicability of device simulators based on drift-diffusion approach (DESSIS and nanoMOS) for InSb MOSFET design and analysis.
3. assess the quality of the device, by analyzing I_{on}/I_{off} ratio, I-V characteristics, threshold voltage, drain induced barrier lowering (DIBL) and unity gain frequency for different configurations and gate lengths using drift-diffusion simulation
4. Study quantum mechanical effects in InSb MOSFET and their effect on transport properties and threshold voltage using drift-diffusion simulation in nanoMOS device simulation program.
5. Investigate scaling rules to remedy possible short channel effects.

Conclusions reached in accomplishing these tasks help us not only to identify areas of performance in which InSb MOSFET will show benefit over other types of transistors but also to point out areas of potential trouble.

2.2 Contributions

1. **Task**: characterize electrical properties of InSb, which has been only partially done in the current literature

   Conclusion and original contribution: developed appropriate and accurate physical models for InSb: electron concentration, density of states, recombination-generation, and mobility. We have developed novel mathematical models to describe 3D and 2D nonparabolic electron concentration which match measured data. These models were successfully implemented in DESSIS and nanoMOS (Matlab). Also, we have successfully implemented Auger model and resolved issues with numerical instability.

   Status: Published in [15, 16]

2. **Task:** Demonstrate applicability of the drift-diffusion simulator (DESSIS) for InSb diodes and MOSFETs

   Conclusion and original contribution: We have shown that the results obtained using models in the task 1 can be implemented in the currently available commercial simulators. We have characterized behavior of InSb diodes and MOSFETs, like differential resistance and I-V characteristics and showed that they match well measured data

   Status: Published in [8] and [15]

3. **Task:** Assess quality of InSb MOSFET device, analyzing I_{on}/I_{off} ratio, DIBL, I-V characteristics, unity gain frequency for different configurations and gate lengths using drift-diffusion simulation

   Conclusion and original contribution: Developed appropriate scaling rules for InSb MOSFET and demonstrated usefulness (as defined above) of the device down to 150nm gate length.

   Status: Published in [8].

4. **Task:** Study quantum mechanical effects in InSb MOSFET on the transport properties and threshold voltage using drift-diffusion simulation

   Conclusion and original contribution: Quantization effects in InSb MOSEFT have not been investigated in the literature and are critical for MOSFET operation. Using an analytic solution we have first demonstrated that nonparabolicity in InSb will have substantial impact on the energy states for narrow quantum wells and cannot be neglected. Therefore, we have developed mathematical model for self-consistent calculation of the effective mass and energy eigenvalues (Schrödinger equation). So far in literature, this was done separately.

   Status: To be submitted for publication [17]

   5. Task: Analyze performance of the dual gate MOSFET and assess impact of InSb nonparabolicity on the performance. Also, identify any

major roadblocks for the further scaling of InSb devices through nanoMOS drift-diffusion simulation.

Original contribution: We have demonstrated that dual gate MOSFET device can be successfully simulated in nanoMOS drift-difusion and quantum mechanical simulator. We show that nonparabolicity affects threshold voltage substantially and has to be taken into consideration. Impact of nonparabolicity on the subthreshold slope and the shape of the electron concentration profile in the channel is small. We have identified mixing of the inversion layers from two gates as a major factor in the reduction of the unity gain frequency, hence suitable scaling rules need to be developed.

Status: To be submitted for publication in [17]

To perform these tasks we used two different simulators: DESSIS and nanoMOS. DESSIS is a commercially available simulator [18], very accurate for silicon device simulation and widely used in industry. However, source code is not open and it has limited options for modification through so-called pmi interface (C-code interface). DESSIS is not designed for highly degenerate and non-parabolic materials like InSb, so creative ways are devised to implement transport properties of InSb. Lack of reliable electrical parameters and poor characterization of InSb makes this task even more difficult.

It has been determined that non-parabolic quantum mechanical treatment cannot be successfully implemented in DESSIS. While DESSIS is more accurate for the calculation of the leakage currents, this is not critical for the cases that require quantum mechanical treatment, due to small volume of the active region. Hence, we used nanoMOS [19, 48] which is an open-source Matlab code developed by Purdue University for simulation of dual gate MOSFET. Since it does not include any recombination-generation mechanisms and it has limited number of physical models for various parameters, such as mobility, it is less accurate then DESSIS, but it can be relatively easily modified to include nonparabolic effects for both transport and quantum mechanical equations. It is primarily used to investigate trends and indentify potential roadblocks in device design. Code is stored on the ICDT website at http://web.cecs.pdx.edu/~hfdmlab/Edin.

In the following section we describe most important electrical parameters of inSb.

3. Material properties

In our simulations we use so called drift-diffusion simulation mode. It is isothermal simulation, suitable for low power density devices which have long active regions. Reason we have chosen drift diffusion over some other alternatives, like hydrodynamic, ballistic or Monte Carlo simulation, is that for the purposes of the computation of the I_{on} and I_{off} it has better numerical stability. Also, size and accuracy of our simulated devices is large enough to make assumptions for drift diffusion to be valid. Basic equations solved are Poisson (eq. (5)), electron and hole continuity equations (6) and current density equations (7) [18]

$$\nabla\varepsilon\cdot\nabla\psi=-q\left(p-n+N_{D^+}-N_{A^-}\right) \tag{5}$$

$$\nabla\vec{J}_n=qR+q\frac{\partial n}{\partial t}$$

$$\nabla\vec{J}_p=-qR-q\frac{\partial p}{\partial t} \tag{6}$$

$$\vec{J}_n=-nq\mu_n\nabla\phi_n$$

$$\vec{J}_p=-pq\mu_p\nabla\phi_p \tag{7}$$

$$n=n_{i,eff}\cdot\exp\left(\frac{-q\left(\phi_n-\psi\right)}{kT}\right)$$

$$p=n_{i,eff}\cdot\exp\left(\frac{q\left(\phi_p-\psi\right)}{kT}\right)$$

Where ψ is electrostatic potential, ε is the electrical permittivity, q is the elementary electronic charge, n and p are the electron and hole densities, and N_D^+ is the number of ionized donors, and N_A^- is the number of ionized acceptors, $n_{i,eff}$ effective intrinsic concentration, R is the net electron–hole recombination rate, μ_n and μ_p are the electron and hole mobilities and ϕ_n and ϕ_p are the electron and hole quasi-Fermi potentials, respectively [18]. These equations are discretized on a given geometry, and we need to know doping and the material properties at each point. Proper models for R and μ_n and μ_p have to be provided.

Implicit in the eq. (5) through (7) is Einstein's relationship between μ and D,

i.e. $\frac{D}{\mu} = \frac{d\eta}{d(\ln(n))}$

where D is diffusivity and η is Fermi level.

3.1 Band structure and density of states

InSb is a direct bandgap material with the smallest bandgap of the III-V material family (E_g=0.17V). The band structure for this material is shown in Fig. 4 . InSb is direct bandgap material, i.e. Γ valley is directly above the peak of the valence band. Next valley is L-valley, 0.68 eV above valence band. This means that bottom of the Γ valley and the bottom of the L-valley are separated by about 0.5 eV at the room temperature. As we will see later, highly doped n-type

InSb (~10^{20} cm^{-3}) will have transfer of electrons from the Γ valley to the L-valley.

Valence band is doubly degenerate and split off band is separated by 0.8 eV from other two valence bands. Split off band will have impact on non parabolicity and cannot be neglected in the calculation of the effective mass, as will be shown later.

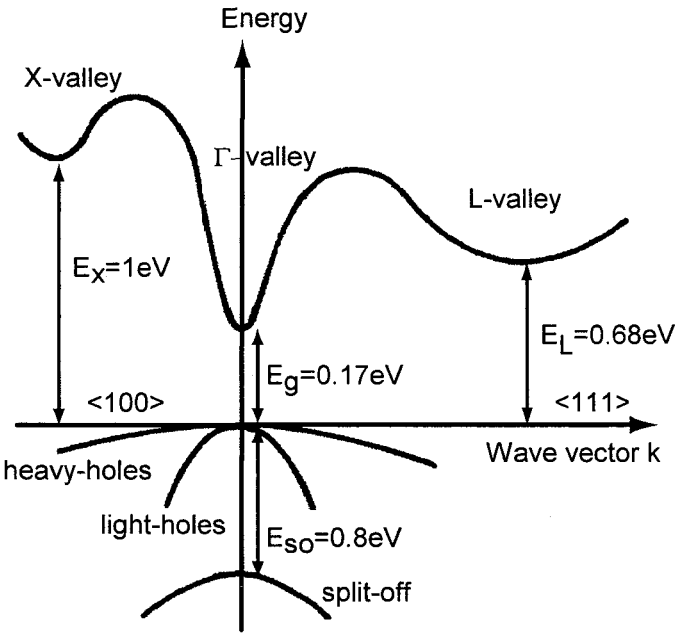

Figure 4: InSb band structure [20]. Copyright by World Scientific

In most commonly used semiconductor materials a standard parabolic band approximation can be used. For some of the III-V compounds this approximation begins to break down and the standard parabolic model is modified to include the non-parabolicity factors, so that the E-k relationship is given by:

$$E(1+\alpha E+\beta E^2+...)=\frac{\hbar^2 k^2}{2m_e} \tag{8}$$

where m_e is the electron effective mass, E is energy, k is wave-vector and α, β are nonparabolicity factors. Usually eq. (8) is truncated to include only the lowest order of nonparabolicity, i.e. just the term containing α.

InSb has a very large non-parabolicity factor of around 5 [1/eV] [20], compared with Si at 0.5 and GaAs at 0.64. Van Wood [21] developed equations to calculate electron concentration for non parabolic materials, but the result was not validated for InSb and final equation is not suitable to be implemented in DESSIS. Smith and Brennan propose dispersion equation but it does not fit experimental data for InSb well [22]. Here we develop independent approach.

To calculate density of states, we find number of states N at the certain wavevector k, contained within [k, k+dk].

$$g(E)=\frac{1}{L^3}\frac{dN}{dE}$$

$$N=\frac{k^3L^3}{3\pi^2}$$

where L has units of real space.

Based on eq. (8) an exact 3D non-parabolic density of states can be derived:

$$g(E)=\frac{\sqrt{2}}{\hbar^2\pi^2}\sqrt{E}m_e^{\frac{3}{2}}\sqrt{1+\alpha E}(1+2\alpha E) \tag{9}$$

which for the limit of small α results in the well known density of states function [23]. Eq. (9) needs to be put into a more suitable form for later integration in eq. (11). Two simplifications were used:

1. only the first term of Taylor's expansion of the $\sqrt{1+\alpha E}$ is retained and it is assumed that $F_{3/2} \approx F_{1/2}$,
2. more accurate approximation is obtained by retaining the first two terms of the Taylor's expansion of $\sqrt{1+\alpha E}$ and neglecting higher terms, so that:

$$g(E) \approx const.\sqrt{E}(1+\tfrac{5}{2}\alpha E - \tfrac{7}{8}\alpha^2 E^2) \qquad (10)$$

Electron concentration is then calculated from:

$$n = \int g(E) f(E) dE \qquad (11)$$

which for the two cases is evaluated to be:

1. $n = \overline{N_C} F_{1/2}(\eta)$,

$$\overline{N_C} = (1+15\frac{\alpha m_e kT}{2\hbar^2})N_C \qquad (12)$$

2. $$n = N_c\{F_{1/2}(\eta) + \frac{15\alpha}{4}F_{3/2}(\eta) + \frac{105\alpha^2 (kT)^2}{32}F_{5/2}(\eta)\} \qquad (13)$$

where $N_C = 2(\frac{m_e kT}{2\pi\hbar^2})^{\frac{3}{2}}$ and $F_{1/2}, F_{3/2}, F_{5/2}$ are the Fermi integrals of order 1/2, 3/2 and 5/2, respectively, with $\eta=(E_F-E_c)/kT$. Using parabolic approximation

electron concentration has the same form as eq. (12) but $\overline{N_c}$ is replaced by the familiar effective density of states N_c. Our results indicate that in the case of

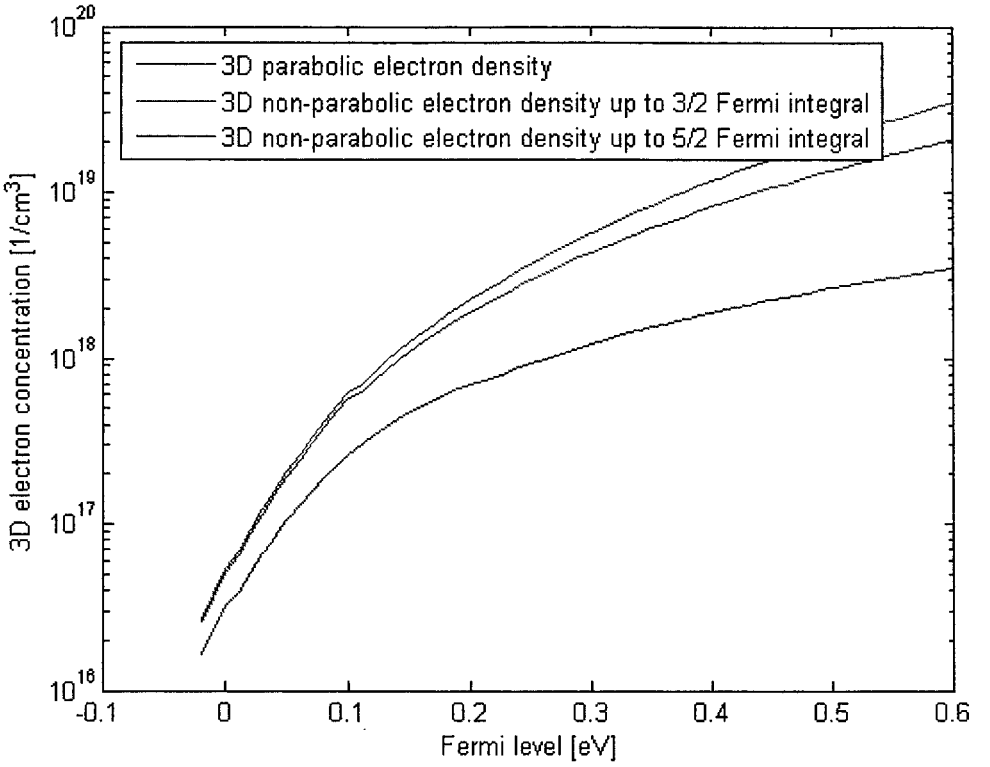

Figure 5: Electron concentration vs. Fermi level for three different approximations for density of states and Fermi integrals (parabolic, eq. (13) truncated at Fermi integral of 3/2 order and full eq. (13)) show significance of nonparabolicity.

eq. (13) integrals of order higher than 5/2 do not affect the results appreciably. Plot of different approximations is shown in Fig. 5. Standard parabolic formula (blue line in Fig. 5) substantially underestimates electron concentration, by about an order of magnitude for Fermi level 0.5 eV above E_c. Error is significant

even for low doping. Since the highest n-type doping in our device is about 10^{19}, parabolic expression for n would result in the Fermi level being higher that the work function, clearly a non-physical result.

Green plot represents eq. (13) truncated at Fermi integral 3/2, while red one is full eq. (13). Small glitch at about 0.1 eV is due to different algorithms used to compute higher order Fermi integrals, depending on whether Fermi level is high or low and it should be neglected. At low doping conditions (up to 10^{18} cm^{-3}), contribution of Fermi integral of the order 5/2 is negligible, but at the highest doping it can be as much as factor of 2. Since 3D electron concentration is used in DESSIS to compute Auger recombination, as will be shown in the section 3.5, accurate treatment is necessary. Therefore, full eq. (13) has to be implemented in the simulator.

3.2 2-D electron concentration

In this section we develop approximate 2D electron concentration formula for nonparabolic materials. 2D expression is needed because nanoMOS calculates 2D densities due to discretization of energy in one dimension. This formula is implemented in nanoMOS.

Two-dimensional electron concentration is in general given by:

$$n = \int D(E) \cdot f(E) dE \tag{14}$$

$$D(E)dE = \frac{k}{\pi} dk \tag{15}$$

$$k = \sqrt{\frac{2 \cdot m^*}{\hbar^2}} \sqrt{\left(E + \alpha E^2\right)} \Rightarrow dk = \sqrt{\frac{2 \cdot m^*}{\hbar^2}} \frac{1 + 2\alpha E}{2\sqrt{\left(E + \alpha E^2\right)}} dE \tag{16}$$

Where D is energy dependent two dimensional density of states, f(E) is Fermi-Dirac probability function and α is non-parabolicity parameter. Substituting k and dk into the expression for the density of states we get:

$$D(E)dE = \frac{1}{\pi} \sqrt{\frac{2 \cdot m^*}{\hbar^2}} \sqrt{\left(E + \alpha E^2\right)} \cdot \sqrt{\frac{2 \cdot m^*}{\hbar^2}} \frac{1 + 2\alpha E}{2\sqrt{\left(E + \alpha E^2\right)}} dE$$

$$D(E)dE = \frac{m^*}{\pi \cdot \hbar^2} (1 + 2\alpha E) dE \tag{17}$$

$$n = \int D(E) \cdot f(E) dE = \frac{m^*}{\pi \cdot \hbar^2} \int \frac{(1 + 2\alpha E)}{1 + e^{(E - Ef)/kT}} dE =$$

$$= \frac{m^*}{\pi \cdot \hbar^2} \left[kTF_0(\eta) + 2 \cdot \alpha \cdot (kT)^2 F_1(\eta) \right] \tag{18}$$

Where F_0 and F_1 are Fermi-Dirac integrals of order zero and one, respectively. Eq. (18) is expression for two-dimensional non-parabolic electron concentration. Note that this expression is exact up to E-k expansion

in α, i.e. up to accuracy of truncating eq. (8) after α. Although eq. (18) is exact it is difficult to implement in a device simulator, due to two Fermi integrals, of order zero and order one. Additional difficulty in using this equation is in the derivation of the Einstein relationship for non-parabolic material, since it involves calculation of the derivative of electron concentration n with the respect to the Fermi level. Here we propose alternative expression. We introduce quantity $\bar{N}_c^*$ given by:

$$\bar{N}_c^* = \frac{m_e kT}{\pi\hbar^2} + \frac{m_e}{\pi\hbar^2} \cdot 2\alpha(kT)^2 \tag{19}$$

This is just a sum of prefactors to the Fermi integrals in the eq. (18). Then we claim that eq. (15) can be approximated by:

$$n = \bar{N}_c F_{1/2}(\eta) \tag{20}$$

Eq. (18) and (20) are compared In Fig. 6 , together with the parabolic expression. Parabolic expression would substantially underestimate electron concentration for a given Fermi level. As can be seen, eq. (20) approximates eq. (18) fairly well: in the worst case electron concentrations differ by 27 % for a given E_F, with an average value around 19%. Since in nanoMOS, which uses these expressions, we do not compute any recombination/generation, this level error in electron concentration can be tolerated. We have implemented eq. (20) in nanoMOS for modeling of dual-gate InSb MOSFET transistors.

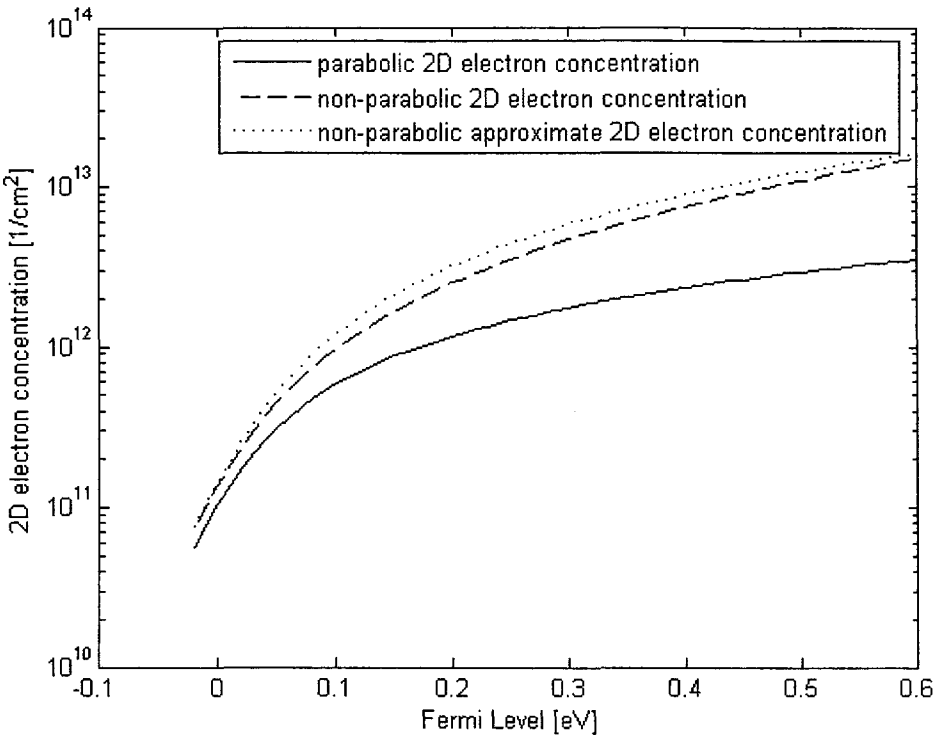

Figure 6: Comparison of different models for 2D electron concentration in InSb

3.3 Degeneracy and intrinsic concentration

Intrinsic concentration is a material parameter that often shows up in modeling of various material properties or processes in semiconductors. InSb, due to its very low bandgap and degeneracy, presents significant challenges in calculating and using this parameter. A material is considered to be degenerate if the Fermi level is less than 3kT (Boltzmann constant x temperature) below the conduction band or above the valence band. This is the point where the exponential Boltzmann approximation and the $F_{1/2}$ integral

start to significantly diverge. To get a rough estimate for the InSb case, we can use:

$$E_i = \frac{E_c+E_v}{2} + \frac{3}{4}kT\ln(\frac{m_h}{m_e}) \rightarrow E_i\text{=0.15eV for InSb} \quad (21)$$

This means E_i is only 0.02eV below the conduction band, well within the 3kT/q (=0.075eV) limit for degeneracy at 300K. Therefore, Fermi-Dirac statistics must be used for electrons in n-type InSb. Note, however, that E_i is far away from the valence band so that Maxwell-Boltzmann statistics can be used for holes in heavily n-doped regions. Fig. 5 illustrates the extreme degeneracy that occurs in n-type InSb, e.g., as doping concentration increases to 10^{19} the Fermi level rises to ~0.45 eV above the conduction band.

Fermi-Dirac statistics will increase computational difficulty of some other models, like quantized energy or Auger recombination-generation. Some other models, like Einstein relationship between mobility and diffusivity, will have to be modified to accommodate Fermi-Dirac statistics [24]. Because of its very small bandgap (E_g = 0.17eV), intrinsic concentration of InSb is very large, e.g. $n_i \approx 2\times10^{16}cm^{-3}$ at 300K, which has traditionally limited its usage to cooled infrared detectors. At 77K n_i drops to ~10^9cm^{-3} which is comparable to Si [$1.6\times10^{10}cm^{-3}$] and GaAs [$1.1\times10^7cm^{-3}$] at 300K. Table 3 below shows intrinsic concentration for different semiconductors at 77K and 300K.

intrinsic concentration [cm^{-3}]	Si	InSb	InAs
T=300 K	10^{10}	$2x10^{16}$	10^{15}
T= 77K	$<10^{3}$	$8x10^{9}$	$5x10^{6}$

Table 3: Intrinsic concentration for different materials at 77K and 300K

Such a high intrinsic concentration could potentially completely dominate any lowly doped InSb regions, thereby necessitating either expensive cooling techniques or carrier extraction-exclusion techniques, as described later in the section 5. Also, intrinsic concentration will appear explicitly in the computation of the Auger recombination-generation rate, so accurate treatment is necessary if the results for, e.g. leakage currents are to be accurate. Plot of the intrinsic concentration, or rather np product in equilibrium, is shown in the Fig. 7 below.

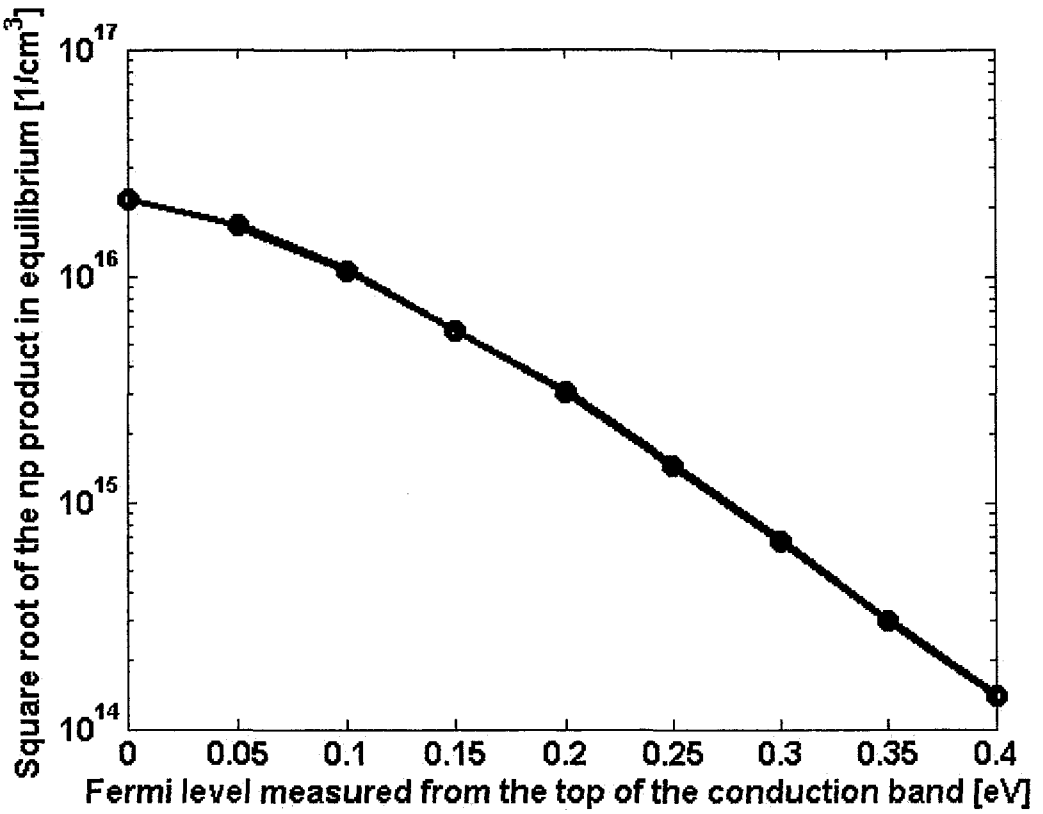

Figure 7: np product in equilibrium vs. Fermi level shows a large variation of two decades.

As can be seen, np product will vary over two decades for different Fermi levels. Doping varies from intrinsic on the left to about 10^{19} cm^{-3} on the right of the graph. Its accurate treatment in the Auger model is a must.

3.4 Mobility and velocity saturation

Intrinsic bulk mobility for InSb is reported to be around $7.8x10^4 cm^2V^{-1}s^{-1}$ [20], which is the largest value of all bulk semiconductors. As expected, the values are much reduced as doping is increased, but the value

does not fall below 4000 even at the highest doping. Intrinsic hole mobility is also relatively high at 800 $cm^2V^{-1}s^{-1}$ and falls to 150 for high doping.

In addition, InSb shows the highest electron saturation velocity, approaching $5x10^7$ cm/s. Combined with very high mobility, it is expected that electron transport in InSb is extremely fast, enabling the design of very fast devices. Note that parasitic resistances in devices would benefit from very high mobilities, i.e. they would be much smaller than in other semiconductors. Parasitic resistances can dominate device performance in deeply scaled devices and this property is not always appreciated in discussions of MOSFET scaling in the literature.

Also, mobility in the channel will be reduced compared to its bulk value. This is due to the surface scattering, which is difficult to model even in silicon. Due to the lack of adequate measurements for InSb, we adjust mobility model for the vertical field dependence to achieve saturation velocity, as described later. However, in quantum mechanical regime, mobility will be increased since the peak of the electron concentration is moved away from the surface, which should reduce surface scattering [25].

In general, mobility data for InSb at 300K is lacking, especially universal mobility model which is widely used for silicon.

3.5 Auger recombination-generation and impact ionization

Leakage current in the InSb diodes is determined by Auger recombination/generation mechanism, in particular Auger-1 and Auger-7 processes [26]. It is a three particle process, where, for example, an electron recombines with a heavy hole releasing energy to another electron (recombination) or high-energy electron impact ionizes an electron–heavy hole pair (generation). Therefore, impact ionization is an inverse process of Auger recombination. Auger-1 is electron initiated process. Therefore it will be more pronounced in n-type material. Auger-7 process is initiated by a hole, making it dominant Auger process in p-type material. Auger processes are important for all low-bandgap materials where activation energy is low. For Auger-1 process, for example, activation energy is the energy required for an electron in valence band to jump to the available state in the conduction band. If the bandgap is small this energy will be small, so we expect leakage currents in these devices to be large. Illustration of the impact ionization is shown in the Fig 8. Electron in the position 1 will generate electron hole pair, denoted by transition from 3 to 4, and loose some energy going to the position 2.

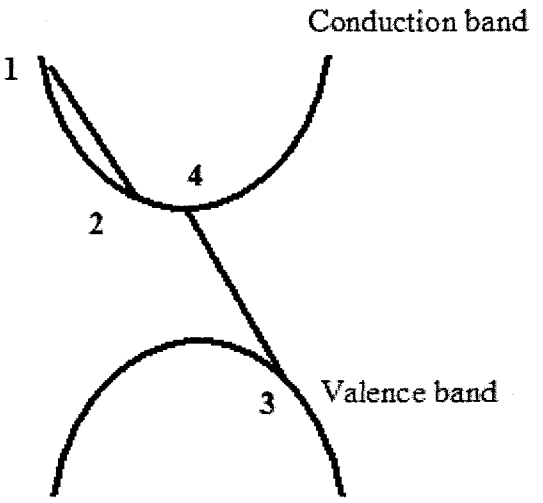

Figure 8: Impact ionization; electron-hole pair is generated contributing to the leakage current in the reverse bias

There has been an effort in the literature to model impact ionization for III-V compounds. Cao and Lei developed formalism to calculate generation rate for GaN [27]. However, GaN has a large bandgap (3.5 eV). Wang and Lei investigated impact ionization for InSb within the framework of the balance equation [28]. These calculations were done at 77K and assuming parabolic bands. Non-parabolicity effects in the Auger generation have been only recently investigated [46, 47]. In our work, we adopt approach from Beattie and White [29]. They proposed following formula for the net recombination rate:

$$R = R_R \cdot \left(1 - e^{-(E_F + H_F + E_g)/kT}\right) \qquad (22)$$

where R_R is recombination in the equilibrium, calculated as a function of the Fermi level E_F which is measured from the bottom of the conduction band and is positive upwards in E-k diagram and H_F is the hole Fermi level measured from the top of the valence band and is positive downwards. In equilibrium, electron and hole Fermi levels align, yielding zero net recombination, due to the exponential factor adding to zero. However, eq. (22) contains an exponential function which causes convergence problems. We were able to successfully implement it for diodes but full MOSFET simulation could not converge. Hence, we had to modify eq. (22) to enable application for InSb MOSFETs.

An alternative expression for Auger R-G is given by:

$$R = (C_n n + C_p p)(np - n_{ie}^2) \qquad (23)$$

where C_n is Auger-1 coefficient and C_p is for Auger-7. While the value of C_p is constant with respect to doping and equal to $5 \cdot 10^{-26} cm^{-6} s^{-1}$, C_n is strong function of electron concentration, due to high degeneracy and non-parabolicity. Plot of C_n dependence on doping is shown in the Fig. 9 below.

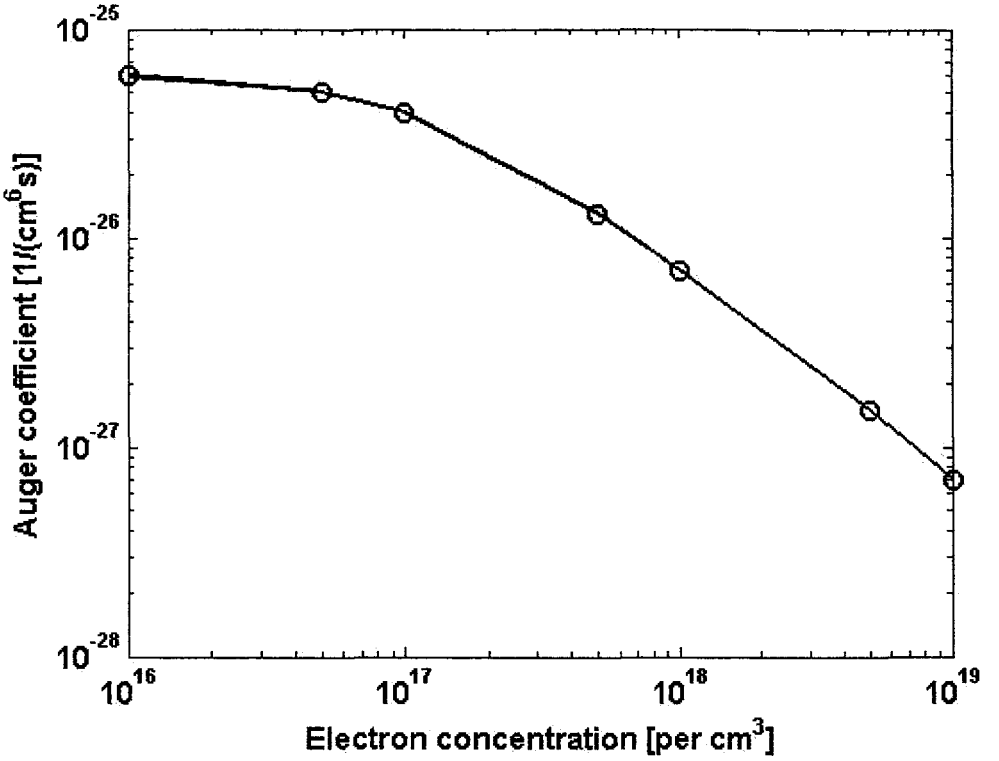

Figure 9: Auger coefficient C_n for electrons in InSb as a function of electron concentration. It is decreasing with increasing concentration due to effective enlargement of the bandgap

This behavior can be explained as follows: as Fermi level moves higher above the bottom of the conduction band there are more electrons available for Auger-1 process, but at the same time energy required to initiate it becomes larger. The latter effect dominates resulting in reduction of C_n as shown in Fig. 9.

Total equilibrium recombination can be written as:

$$R_R = R_{A-1} + R_{A-7} + R_{SRH} \qquad (24)$$

Three terms indicate equilibrium recombination due to Auger-1, Auger-7 and Shockley–Reed–Hall process (SRH), respectively. Beattie and White evaluated equilibrium recombination for Auger-1 at four different temperatures: 294 K, 220 K, 150 K and 80 K. For temperatures between these points, we used linear interpolation. Auger recombination peaks at some value of E_F above the bottom of the conduction band. It rapidly decreases as Fermi level increases above the peak value, due to strong degeneracy (increase in the activation energy). This is illustrated in the Fig. 10. As material becomes more p-type due to lack of electrons to initiate Auger-1, Auger process decreases.

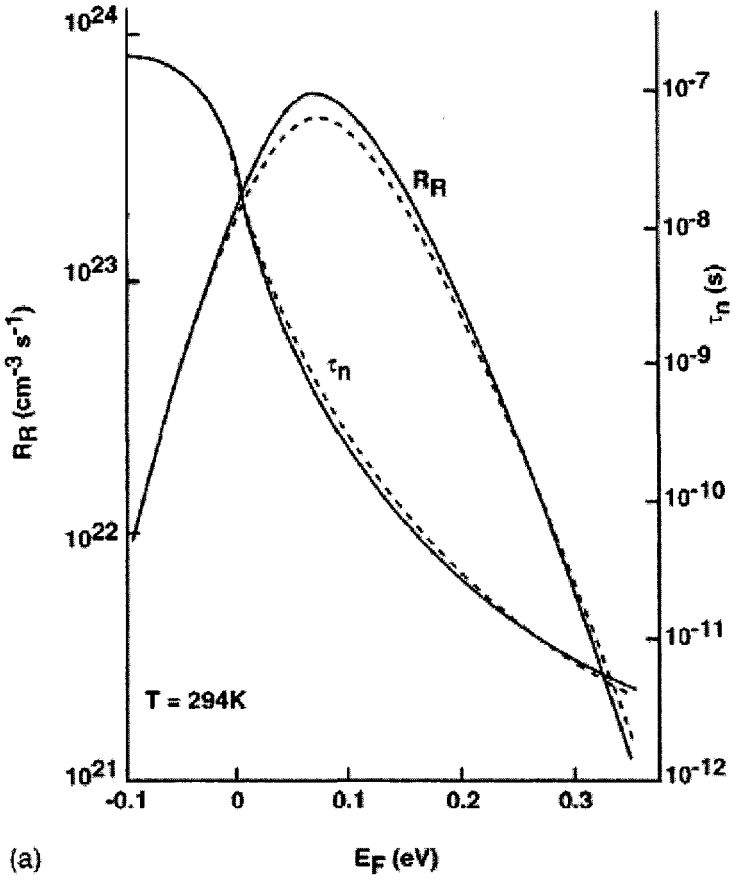

Figure 10: Equilibrium Auger recombination and electron lifetime vs Fermi level: solid lines are from full band calculation, dashed from Beattie model in eq. 22 and 24. It reaches a maximum slightly above the bottom of the conduction band. Copyright by the American Institute of Physics

Equilibrium recombination rate for A-7 process is approximately 3 times smaller than for Auger-1 [30]. Therefore, in our implementation of the Auger-7 process we are simply using Auger-1 values divided by 3.

3.6 Shockley-Reed-Hall (SRH) recombination

Using the exclusion/extraction method described in the following section, Auger processes in the lowly doped π region are suppressed. Under such conditions SRH process becomes a comparable contributor to the overall generation-recombination at 300K. At lower temperatures SRH becomes the dominant process [29, 30, 31]. According to the measurements performed by Nott et al [31] on InSb/InAlSb diode, SRH coefficient in the π region is $A=9.26\times10^{6}\ s^{-1}$. In our devices, this would result in the equilibrium recombination of $R_{SRH}=4.63\times10^{22}\ cm^{-3}s^{-1}$ according to $R_{SRH} = N_T \cdot A$, where N_T is trap density. We assume N_T to be equal to the doping i.e. $N_T \approx 10^{15}\ cm^{-3}$ Unfortunately, no measurements of the SRH coefficient for InSb for non-intrinsic conditions are available in the literature. However, it can be argued that they are not critcial since our MOSFET channel is close to intrinsic and in other regions of the device, Auger process dominates. Also, numbers presented for SRH recombination are somewhat arbitrary, since they depend on the processing and it is generally assumed to be related to how "clean" some process is. We have, therefore, decided to keep the SRH model simple and use a constant coefficient provided by Nott et al.

3.7 Band-to-Band tunneling

Due to narrow bandgap, even at the moderate doping of the p-n junction, one may expect substantial tunneling current. The band-to-band-tunneling (BTBT) is modeled as an R-G process, according to [32]:

$$R_{bbt} = -B|F|^{\sigma} D(F, E, E_{fn}, E_{fp}) \exp(-F_0 / |F|) \quad (25)$$

where B and F_0 are material dependent parameters, F is field, E_{fn} and E_{fp} are quasi-Fermi levels. Their values for InSb are determined according to Kane's model [33]. D is a function that accounts for the relative position of electron and hole Fermi levels in the neutral regions and the influence of perpendicular electron motion on tunneling probability. As will be shown below, BTBT has no influence on leakage current in InSb diodes.

All of the processes: Auger, SRH and BTBT are used in DESSIS but, as noted earlier, all recombination-generation mechanisms are turned off in nanoMOS. In the next section we discuss DESSIS implementation of the InSb electrical properties.

4. DESSIS Implementation

For simulation purposes, the DESSIS device simulator from ISE (now Synopsis) [18] was used. This software allows for custom material definition using at a minimum the following set of basic parameters (for drift-diffusion simulations): ε - the relative dielectric constant; E_g -the bandgap; χ - the electron affinity, eDOS, - the electron density of states mass; hDOS - the hole density of states mass, and μ - the carrier mobility. These parameters were obtained from [20].

E_g = 0.17eV (at 300 K),

χ = 4.59 eV

hDOS = $0.43m_0$,

eDOS = $0.014m_0$

Calculation of density of states for electrons (DOS) in Dessis can be done in two ways:

a) by defining a constant DOS, or

b) by defining a density of states effective mass (mDOS), which is then used to calculate DOS.

Maxwell-Boltzmann (M-B) or Fermi-Dirac (F-D) statistics can be used in either case. If Fermi-Dirac statistics is used, only $F_{1/2}$ integrals are calculated within DESSIS.

Given these options, the only way to implement eq. (13) is by utilizing approach b) and making the mDOS a function of doping, where its value is adjusted empirically to give the appropriate Fermi level for a given doping. This was implemented using C++ "pmi" subroutine. Results shown in Fig. 5 agree with those in [34].

Since heavy hole band is parabolic, no special treatment is needed in DESSIS, i.e. hole concentration is calculated through the standard Fermi-Dirac statistics formula.

Basic set of parameters for mobility model was obtained mostly from Levinshtein [20] and implemented as custom material. In DESSIS, we use Arora model [35, 36] for doping-dependent degradation:

$$\mu_{DOP} = \mu_{MIN} + \frac{\mu_D}{1 + \left(\frac{N_i}{N_0}\right)^{A^*}} \tag{26}$$

, where

$$\mu_{MIN} = A_{MIN} \cdot \left(\frac{T}{T_0}\right)^{\alpha_m} \qquad \mu_D = A_D \cdot \left(\frac{T}{T_0}\right)^{\alpha_D} \tag{27}$$

$$N_0 = A_N \cdot \left(\frac{T}{T_0}\right)^{\alpha_N} \qquad A^* = A_A \cdot \left(\frac{T}{T_0}\right)^{\alpha_a}$$

Individual parameters for InSb are given in the Table 4 and N_i is doping concentration:

	N	p	Units
Ar_mumin	$3x10^3$	200	$cm^2/V/s$
Ar_alm	0	0	1
Ar_mud	$7.4x10^4$	500	$cm^2/V/s$
Ar_ald	-1.538	-2.108	1
Ar_N0	$2.7x10^{17}$	$7x10^{17}$	cm^{-3}
Ar_alN	5	3.45	1
Ar_a	0.9187	0.521	1
Ar_ala	1.1061	0.1077	1

Table 4: Arora mobility model parameters

To model velocity saturation we turn on model 1 in DESSIS [18] which is based on the Canali model [36]:

$$\mu(F) = \frac{\mu_{low}}{\left[1 + \left(\frac{\mu_{low} F}{v_{sat}}\right)^{\beta}\right]^{1/\beta}} \tag{28}$$

where $v_{sat} = 5 \times 10^7 cm/s$ is saturation velocity, mu_low is low field mobility obtained from the Arora model described above, $\beta=2$ and F is longitudinal electric field.

Due to the strong transverse electric field in the channel, there is an interaction between electrons and InSb-insulator boundary. In DESSIS, this effect is implemented using Lombardi model. Parameters are obtained by optimization to achieve proper velocity profile in the channel, i.e. most of the channel

achieves velocity saturation. Reduction of mobility was, in relative terms, the same as for Si . Furthermore, in our simulations of the reference device the maximum transconductance g_m was around 125 mS/mm, which is very close to experimental data of 120 mS/mm [6], indicating a reasonable fit of our mobility model. Further model refinement and verification is not possible without more experimental data. Parameters for the Lombardi model are given in Table 5:

	n	p	Units
B	2.78×10^8	4.93×10^7	cm/s
C	8.8×10^4	8.95×10^5	$cm^{5/3}sV^{2/3}$
N_0	1	1	cm^{-3}
λ	0.125	0.0317	1
k	1	1	1
δ	5.82×10^{15}	2.05×10^{15}	V/s
A	2	2	1
α	0	0	1
N_1	1	1	cm^{-3}
v	1	1	1
η	5.82×10^{32}	2.05×10^{32}	$V^2/cm\ s$
l_crit	1×10^{-6}	1×10^{-6}	cm/s

Table 5: Lombardi model parameters for InSb

Dependence on the vertical field is given in the Fig. 11, to illustrate that mobility can be dramatically reduced by vertical field. In our device, fields can be in excess of 10^6 V/cm.

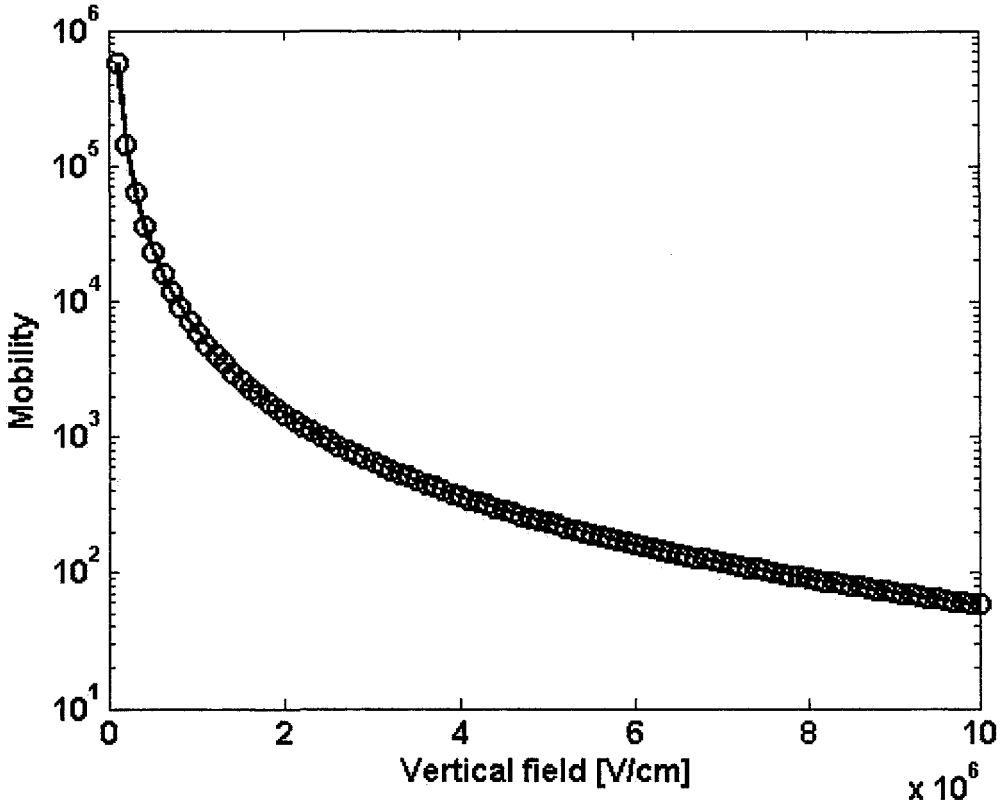

Figure 11: Mobility vs. vertical field plot shows rapid decrease in electron mobility with the vertical field

5. InSb diodes and exclusion/extraction mechanism

Exclusion/extraction principle is a design approach to overcome the high intrinsic concentration of InSb at room temperature and the corresponding leakage currents. High leakage current would reduce I_{on}/I_{off} ratio, rendering the device useless since the off state power would be prohibitively high. In a MOSFET design this approach is utilized to reduce the leakage coming from source-body and drain-body junctions.

The drift diffusion models described here were used to examine the possibilities of InSb p-i-n diodes operating at room temperature. The devices were first published by Ashley, et. al. [37]. The extraction (n^+-π-p^+) and exclusion/extraction (n^+-π-$\underline{p}^+$-p^+) diodes are two different strategies to overcome the high intrinsic concentration of InSb at room temperature and the corresponding leakage currents. Cross sections of both diodes are given in the Fig. 12. Following is a brief description of how these devices operate.

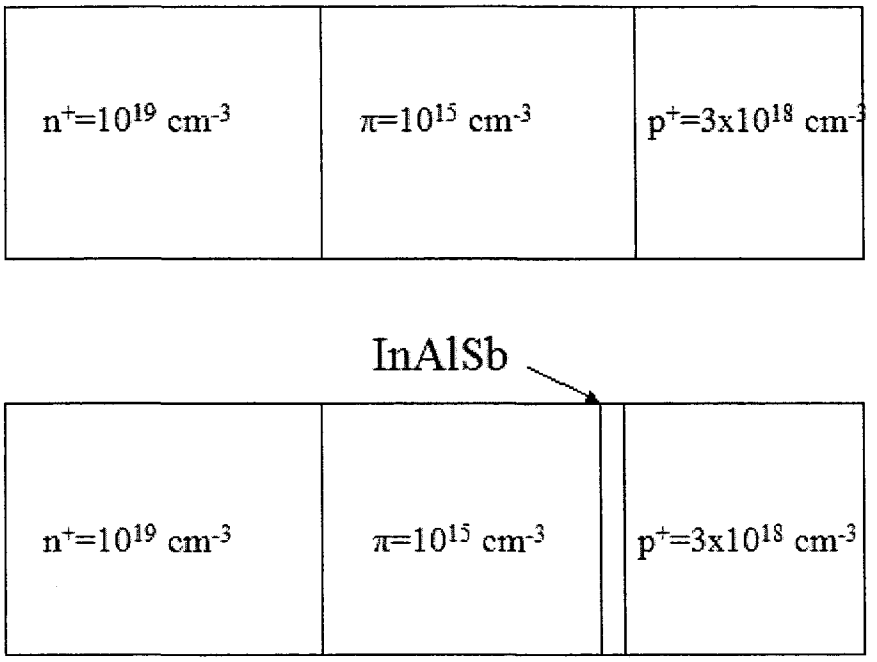

Figure 12: Extraction (top) and exclusion/extraction (bottom) diode cross section

5.1 Extraction Diode

The extraction diode is the simpler of the two and it consists of a heavily doped n^+ layer, followed by a very lowly doped p layer (the π region) and a heavily doped p^+ region yielding an n^+-π-p^+ structure. Due to the extreme degeneracy in heavily doped n-type InSb, the Fermi level moves high into the conduction band so that E_F-E_v =(E_F – E_c) + Eg is 2-3 times larger than the bandgap (see Fig. 5). This large difference drives the concentration of minority holes to very low levels in the n^+ region. As a result, intrinsic concentration n_{ie}

(or, more accurately, np product in equilibrium) is much reduced in highly doped n^+ region. For reverse bias, generation in n^+ region which happens close to the depletion layer depends on n_{ie}^2 (see eq. (23)) and is thus reduced by several orders of magnitude.

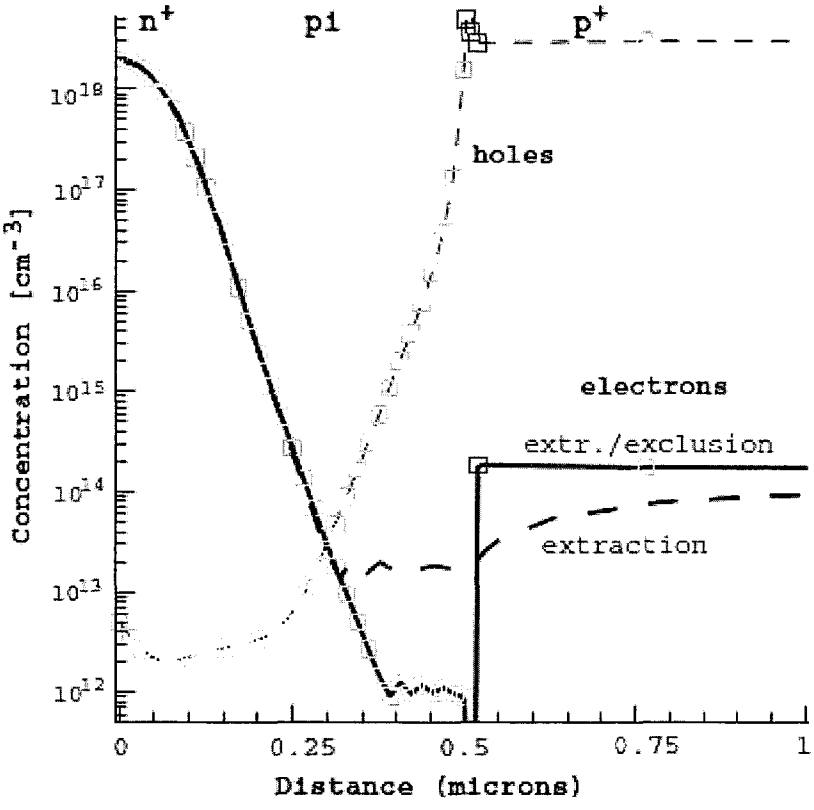

Figure 13: Electron and hole concentrations for extraction and extraction/exclusion diodes under reverse bias V=-0.35 V

Degeneracy exists in p^+ InSb material too, but to a much lesser extent than in n-type material. Due to the high density of states in the heavy hole band, degeneracy is not capable of reducing n_i below its "normal" value for intrinsic material because $E_F - E_v \approx E_g$. In reverse bias there still exists a significant electron concentration that normally diffuses into the π region contributing to the leakage current (see Fig. 13). The minority concentration has an exponential "tail" (long diode case) that extends into neutral p^+ layer and eventually reaches the equilibrium value inside it, some distance away from the depletion layer. Inside this transition layer $np-n_{ie}^2$ term goes from being dominated by n_{ie}^2 (large generation) to zero (i.e. recombination = generation). Therefore, the total current has contributions from diffusion of electrons as well as from generation inside the p^+ region resulting in large leakage currents. Overall effect of the extraction mechanism on the leakage current will be presented in section 6.

5.2 Exclusion/extraction Diode

In order to prevent electrons from entering the π region, a thin region of a wide bandgap material, p^+, is placed between the p^+ and π regions resulting in a n+-π-p+-p+ diode. It needs to be thick enough to prevent tunneling through the barrier and high enough to prevent injection over the barrier, but it cannot be too thick because of the lattice-mismatch strain. A layer of InAlSb with ~10 nanometers thickness was used in [37] and in our simulations. The

bandgap offset is assumed to show up almost entirely in the conduction band. The p^+ region now acts as an "exclusion" layer (or contact), so the diode is called "exclusion/extraction" diode.

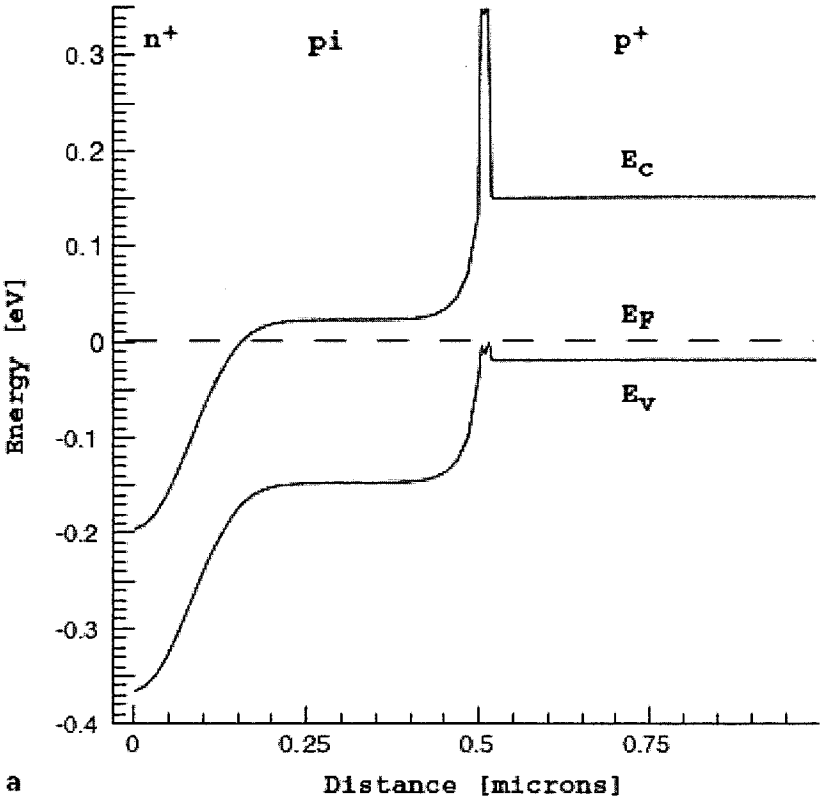

Figure 14: Band diagrams for $n^+ - \pi - p^+ - p^+$ exclusion/extraction diode in equilibrium

Fig. 14 show equilibrium band diagrams for n^+-π-p^+-p^+ diode. Notice the spike in the conduction band corresponding to the insertion of the InAlSb material.

n^+-π-p^+ diode diagrams look very similar, except for the absence of the spike. This spike effectively prevents any injection of minority electrons into p^+ region. Also note the high degeneracy in the n^+ region on the left.

Final illustration of our explanation of operation of these two diodes, is shown in Fig. 15 where the net generation rate is shown. In extraction diode there is a long tail of generation extending into p^+ region, while exclusion/extraction diode has a very narrow region where generation occurs. This peak occurs precisely at the π-p^+ interface and it is bigger for the extraction diode since there is no wide gap layer to prevent injection of the electrons from p^+ region. Clearly, Auger generation has a very large effect on leakage current of extraction diode, but much smaller one on exclusion/extraction diode.

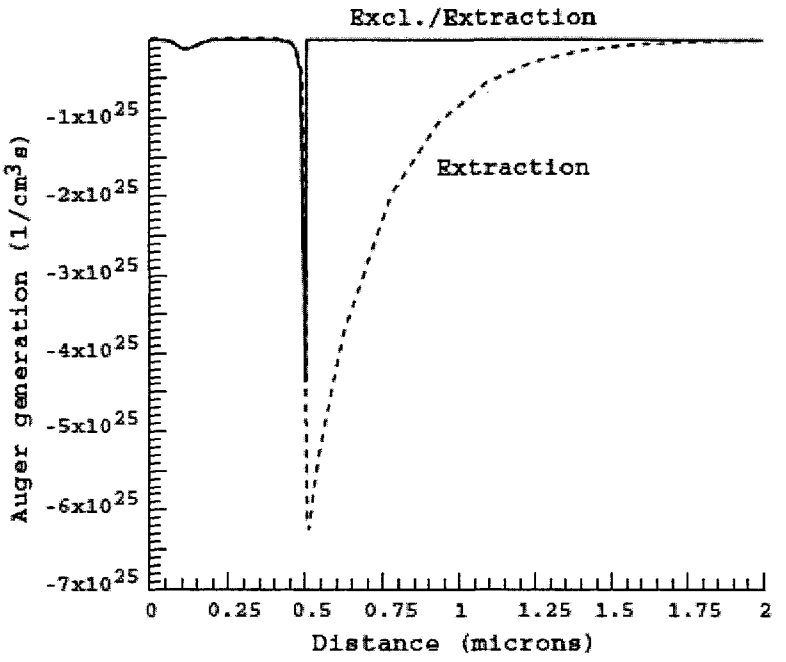

Figure 15: Auger generation in n^+–p–p^+ and $n^+ - \pi - p^+ - p^+$ diodes under reverse bias. Exclusion/extraction diode has reduced generation by more than a decade in p^+ region close to p-n junction.

6. Diode simulations and results

In order to validate our models and verify exclusion/extraction mechanism, we first constructed and simulated two InSb diodes, as described above. Simulation results for electron and hole densities in each region of these devices are shown in Fig. 13 and we observed the following:

a) the difference in electron concentrations inside the π region and

b) sharp, step-like, increase of n for n^+- π-p^+-p^+ diode vs. exponential increase for n^+-π-p^+ diode inside the p^+ region.

In the following sections we discuss how this affects terminal behavior of diodes.

6.1 I-V characteristics

Additional diode I-V simulations were performed to determine the effectiveness of suppression of diffusion current and Auger generation in InSb exclusion/extraction and extraction diodes. One can see in Fig. 16 that the leakage current in the exclusion/extraction diode is more than an order of magnitude smaller than the simpler extraction diode. This is the region on the left of the plot, and since in our MOSFET device body will always be reversely biased, forward bias results can be ignored, but are shown for completeness.

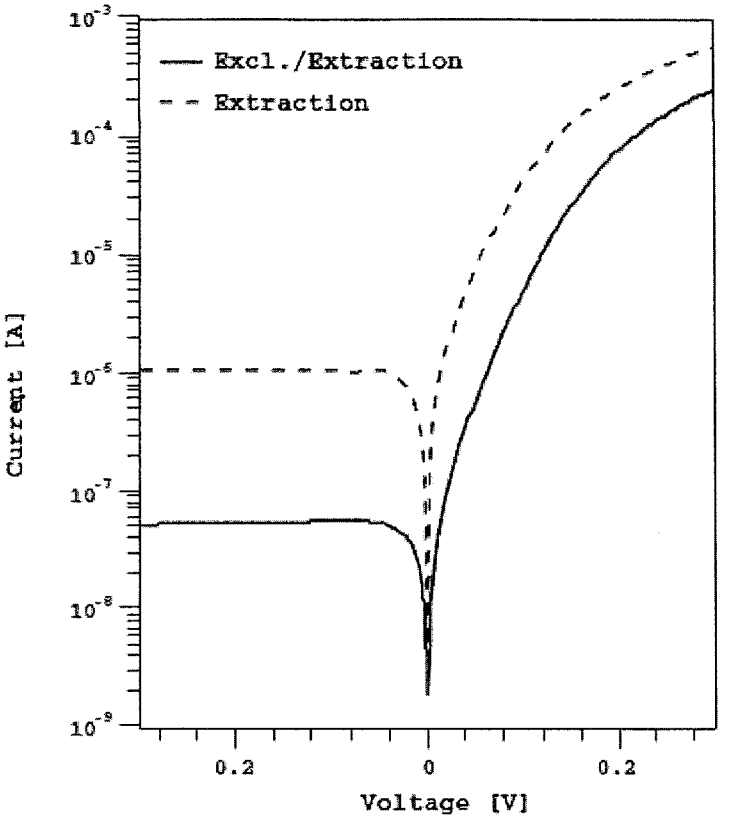

Figure 16: Leakage currents in $n^+ - \pi - p^+ - p^+$ (exclusion/extraction) and n^+-π-p^+ (extraction) diodes show improvements in the reverse bias by addition of thin InAlSb layer

6.2 Differential resistance

Another commonly used method that allows us to evaluate the effectiveness of carrier extraction/exclusion is to calculate the zero bias differential resistance R_0. To accomplish this, a diode voltage is swept from negative to positive and the slope of the curve at 0 V bias is measured. This

measurement is commonly used in photodiodes to compare sensitivity. Higher resistivity results in higher detector sensitivities, *i.e.,* for a given number of electron hole pairs, more voltage is generated in the detector.

An analytical expression for R_0A, where A is cross sectional area of the diode, can be obtained starting with the diode current equation

$I=I_0(exp(V/V_T)-1)$ (29)

If the derivative of the above expression is taken with respect to V one finds:

$$\frac{dI}{dV}=\frac{I_0}{V_T}\exp(\frac{V}{V_T})=\frac{I_0}{V_T}[\exp(\frac{V}{V_T})-1]+\frac{I_0}{V_T}=\frac{I}{V_T}+\frac{I_0}{V_T} \qquad (30)$$

When $V=0$, $I=0$, and the expression for differential conductance reduces to:

$$\frac{dI}{dV}=\frac{I_0}{V_T}=\frac{I_0}{kT/q} \qquad (31)$$

The inverse of differential conductance is the differential resistance, R_0 . so that

$$R_0A=\frac{kT/q}{I_0}A \qquad (32)$$

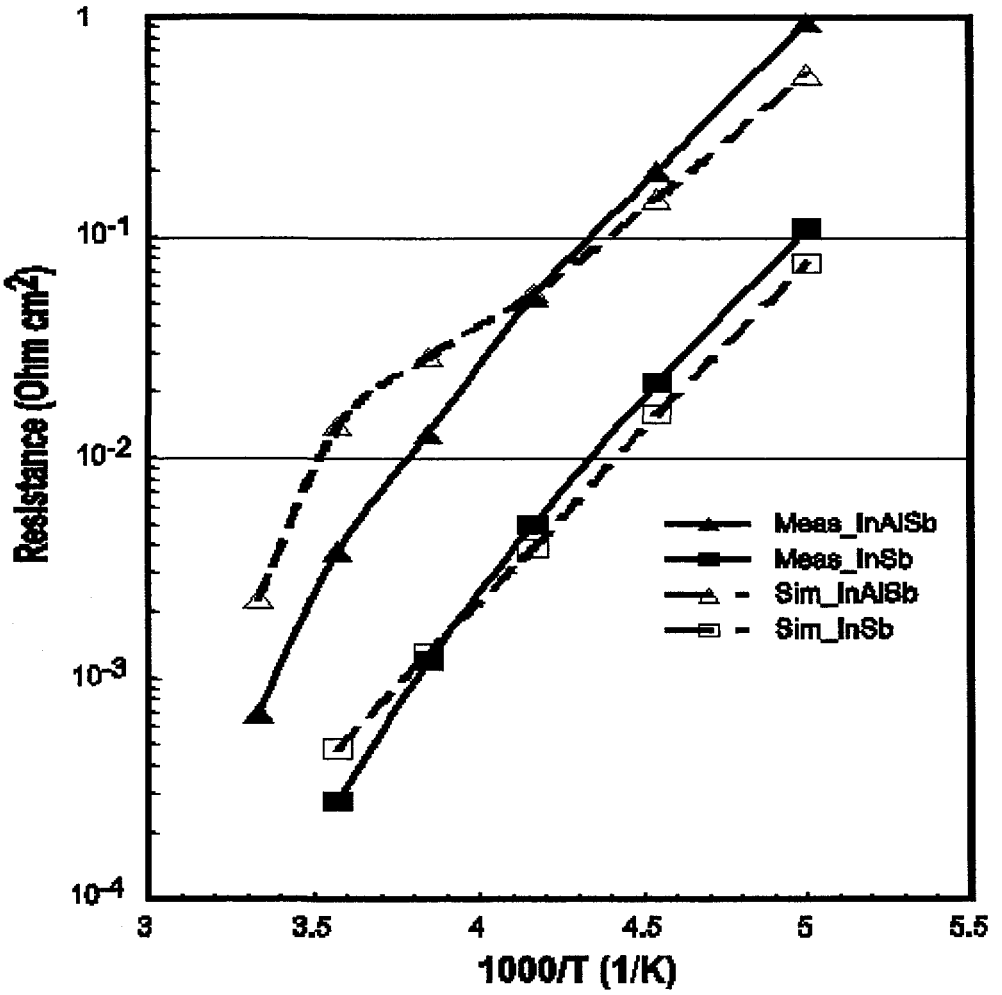

Figure 17: Differential Resistance: simulation vs. measured data for extraction diode (labeled "InSb") and exclusion/extraction diode (labeled" InAlSb"). Extraction diode shows relatively good fit with the measured data.

From eq. (32) it is clear that R_0A is temperature dependent: explicitly thorough T and implicitly through I_0. Since main contributor to the leakage current is Auger generation, which is strongly dependent on temperature, R_0A will depend on T through it. If we plot R_0A as a function of inverse temperature, we

can see the difference in resistivities of the different types of diodes. We note that this difference is fundamentally due to the difference in leakage currents of these diodes, which in turn is related to the differing levels of effectiveness of the extraction/exclusion mechanisms. By measuring the slope of the I_d vs. V_d curve at the 0V Fig. 17 is obtained that matches well with published data [37].

There are two important observations to be made from Fig. 17. First, for the extraction diode we see a good fit between our modeled diode and actual measurements. These curves represent a simple extraction InSb diode without the thin higher bandgap InAlSb layer. As theory indicates, and measurements confirm, recombination and generation are dominated by Auger processes and the good match between simulation and measurement confirms the validity of our methodology. The upper two curves in Fig. 17 represent an exclusion/extraction diode including the InAlSb layer. Based upon the physical arguments advanced previously for exclusion/extraction diodes, we expect that Auger processes should be much smaller in these diodes and that some, or even most, of the carrier generation would come from the SRH mechanism. Unfortunately, there were no data available to accurately determine appropriate SRH parameters. The value chosen for SRH lifetime ($5x10^{-8}$ s) present an optimal fit over the full range of temperatures in Fig. 18. Other SRH values were investigated that yielded better matching at one end or the other of the temperature spectrum, but

always to the detriment of the matching at the opposite temperature extreme. Further experimental data is needed to better comprehend the SRH effect in the InSb/InAlSb case, at which point more consistent results between simulation and measurement should be possible. Turning BTBT model on and off produced no appreciable change in currents, indicating that this mechanism plays no role in InSb devices in the investigated temperature range.

7. Exclusion/extraction MOSFET

In this section we describe InSb MOSFET, as originally proposed by Ashley et al. in [6]. Subsequently, this device design received more attention [8, 37], and application of InSb was further expanded [5] to include heterojunction devices. Although there is a considerable interest in InSb based HEMTs, in our work we focus on MOSFET structures [4]. The main reason for this is that HEMTs have Schottky barrier for gate contact, making this normally-on device, i.e. channel will conduct for zero voltage applied on the gate. Making gate Schottky diode in the enhancement mode is not simple. For digital application we prefer normally off devices, so recently there has been interest in making even GaAs devices as MOSFET structures. MOSFET structure can be easily made in the depletion mode. Furthermore, classic HEMT design has reached a scaling limit at around 100nm and further scaling will require different gate design [11].

7.1 Device description

Idealized cross section for the reference device is in Fig. 18, while Fig. 19 shows the cross section of the InSb MOSFET in DESSIS.

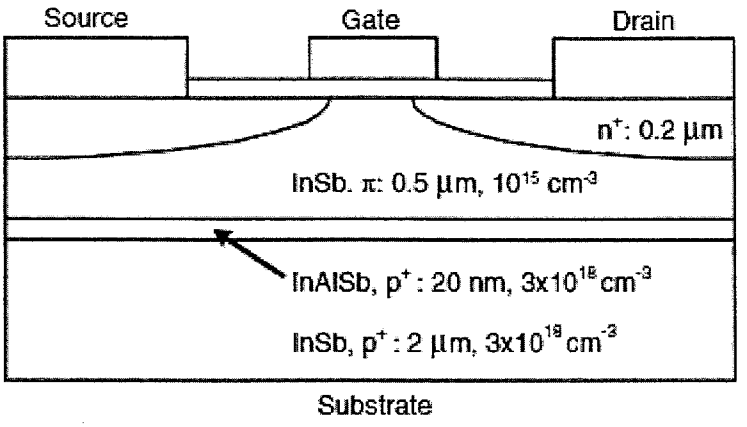

Figure 18: Idealized cross section of the InSb MOSFET based on [6]

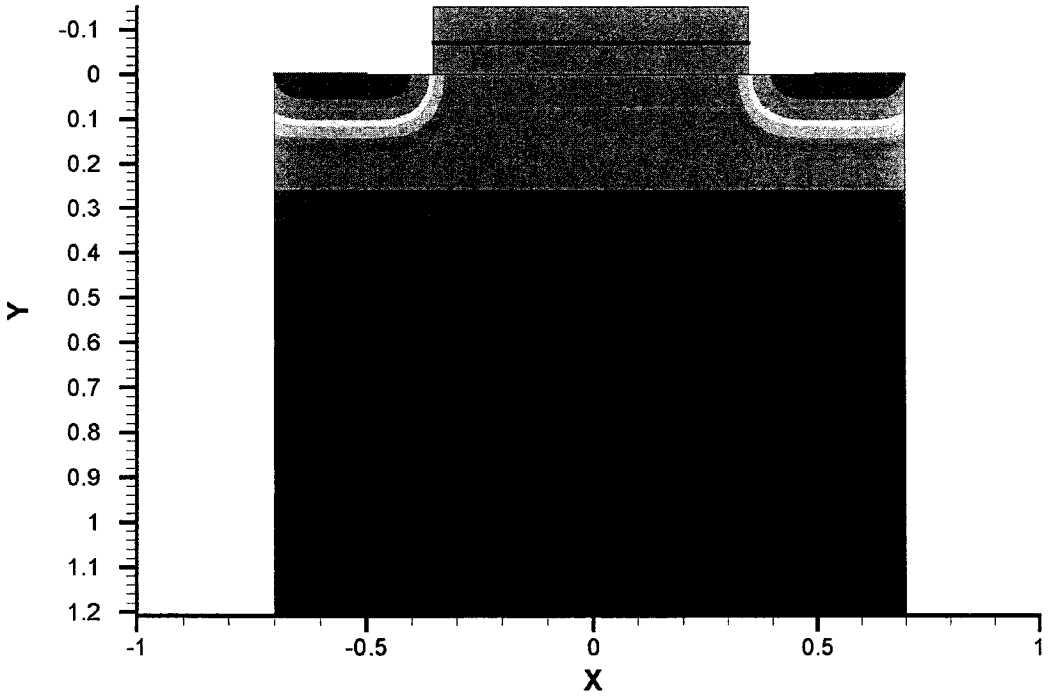

Figure 19: InSb MOSFET cross section, as used in DESSIS all units in µm

Substrate is 2μm thick p^+ InSb layer, with constant doping of $3x10^{18}$ cm^{-3}. On top of it is thin (20 nm) layer of $In_{0.85}Al_{0.15}Sb$, which is also p^+ with doping $3x10^{18}$ cm^{-3}. Active region is 0.5μm thick and at the room temperature it is intrinsic. Background doping is 10^{15} cm^{-3}, p type. Source and drain have a Gaussian profiles with peak concentration of 10^{19} cm^{-3} at the surface. Junction depth is 0.2μm. Gate oxide is 70nm thick SiO_2. Gate, source and drain electrodes are made of gold.

After we varied substrate contact voltage, we found that –0.35V maximizes differential resistance of the source (drain)-substrate diode, thus minimizing leakage. Therefore, this value is fixed for our MOSFET simulations. Due to exclusion/extraction, as explained in the section 6, electron concentration in the π region is reduced well below its intrinsic value, while hole concentration is pinned down to the background acceptor concentration (10^{15} cm^{-3}). This in turn, reduces Auger R-G resulting in reduced leakage currents.

7.2 Short Channel Effects

One of the additional advantages of MOSFET devices is relatively well understood scaling that happened over the past few decades. To put it simply, scaling is the reduction in the size of the MOSFET device. One of the obvious benefits of scaling is ability to put more devices in the same area, hence increasing functionality, or reducing area for the same transistor count, hence

reducing the cost. This ability to scale was the main driver of the Moore's law. Also, reducing size of the transistor reduces gate capacitance and gate length, from which we would expect smaller devices to switch faster.

However, scaling has its problems as well. Reducing size will create myriad of operational problems with manufacturing and reliability. Furthermore, physics of the short-channel devices introduces unwanted effects. Among those, we will focus on threshold voltage reduction, which makes it impossible to switch the device completely off for zero volts on the gate, i.e. V_{gs}=0V. Usually, as a rule of thumb, we require three decades separation between I_{on} and I_{off} to clearly define logic states. I_{off} is defined as the drain-source current at zero volts bias at the gate, while we define I_{on} current as the current with 1 V applied at the gate.

Another short-channel effect is the drain induced barrier lowering (DIBL). Under off conditions, potential barrier in the p-type region will prevent electrons flowing from source to the drain. For a long channel, source and drain fields will not penetrate deep into the channel, making this barrier virtually flat over the channel length. As gate length is reduced, source and drain fields will penetrate more, making the barrier lower. This, in turn, will create larger subthreshold current, reducing threshold voltage compared to the long channel case. If we apply higher drain voltage, drain fields will be even stronger, lowering the barrier further. Therefore, threshold voltage will become dependent on the drain voltage, and increment between different values for

the drain voltage can be rather high, as we will show later [38]. For DIBL to be acceptable, it should not be more than 100 mV per V.

7.3. Simple scaling results

To demonstrate these effects, several gate lengths were simulated: 0.7, 0.5 and 0.3μm. Reference device is 0.7μm, and its threshold voltage is 0.45V. Scaling is simple, i.e. only gate length is reduced by a set factor, while doping, voltages and other dimensions are preserved. In this way, short channel effects are observed. In Fig. 20, I_{ds} vs. V_{gs} for different gate lengths is shown. V_{ds} is held at 0.5V. For 0.7μm I_{on}/I_{off} separation is two and a half decades. However, due to the reduction in the threshold voltage, smaller gate lengths have much worse I_{on}/I_{off} ratio. This suggests that short channel effects become dominant already at 0.3μm gate length, where I_{on}/I_{off} separation is less than a decade. Subthreshold slope for this device is very large.

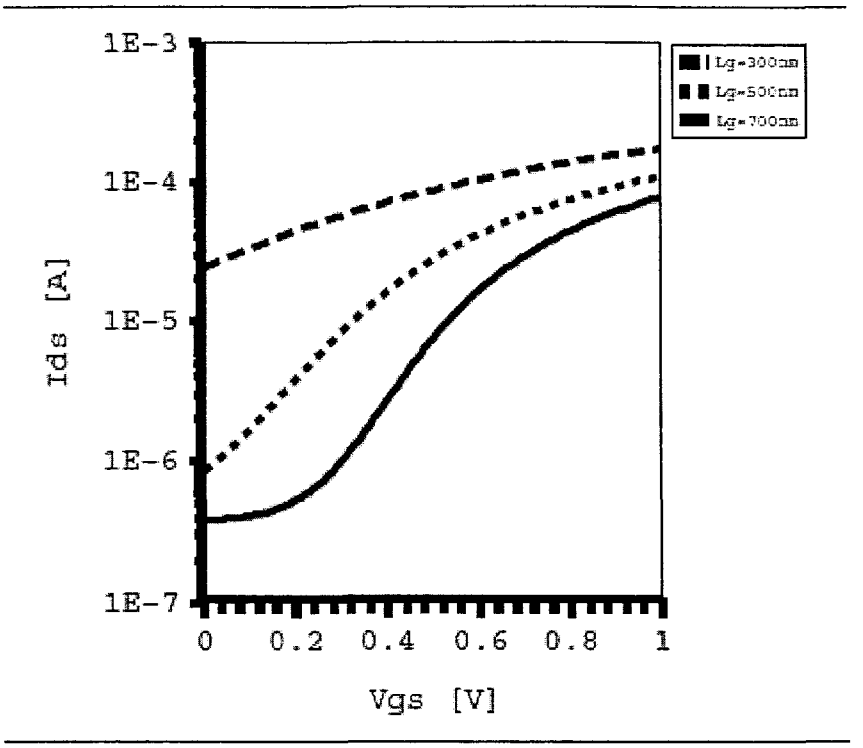

Figure 20: I_{ds} vs V_{gs} for different gate lengths; short channel effects become dominant at 300nm device

The other short channel effect we simulated is the drain induced barrier lowering (DIBL), as explained earlier. We measured threshold voltage by extrapolating I-V curve in the linear regime down to the gate voltage axis. Voltage at which it intersects axis is the threshold voltage V_T. Using this procedure, we measured threshold voltage at two different values of V_{ds}, 0.05V and 0.5V, for each gate length. Threshold voltages are summarized in Table 6.

Gate length [μm]	$V_T(V_{ds}=0.5V)$	$V_T(V_{ds}=0.05V)$
0.7	0.5	0.55
0.5	0.35	0.45
0.3	negative	0.35

Table 6: Threshold voltage [V_T] vs. V_{ds} and L_g

At shorter gate lengths, performance deteriorates and DIBL can cause serious shifts in V_T. For relatively long gate lengths, 0.7 μm, this shift is small, 0.05 V. However, for 0.3 μm device it is more than 0.35V, rendering device unusable. To remedy those short channel effects, proper scaling rules are needed.

7.4. Device scaling results

There are several different scaling strategies employed in the industry. Our approach is modified quasi-constant field scaling. In constant field scaling, horizontal and vertical dimensions of the device are reduced by the same factor, while applied voltage is decreased. Also, substrate doping is increased by this factor. This way electric field will remain unchanged between different device sizes [38].

Due to peculiarities of InSb material and the device design used, these scaling rules need to be modified, as we will explain below. Hence, we call this approach quasi-constant field scaling. As gate length is reduced by the factor k, depth of the $In_{0.85}Al_{0.15}Sb$ layer, junction depth and thickness of the gate

oxide are reduced by the same factor. Also, in order to increase threshold voltage for shorter gate devices, it is necessary to increase the p-type doping in the channel region. This is accomplished by using a shallow Gaussian doping, with junction depth of 43nm. Peak concentration is increased as gate length is reduced, as shown in the Table 7. However, since we do not increase doping in the entire substrate, but just in this shallow region, this is not canonical constant field scaling.

Gate length [um]	N_a (peak)
0.5	$1.8 \cdot 10^{17}$
0.3	$3 \cdot 10^{17}$
0.15	$6 \cdot 10^{17}$

Table 7: Channel doping vs. L_g

Reference device has Lg= 0.7μm, and its threshold voltage is 0.45V. For this device I_{on}/I_{off} separation is two and a half decades (see Fig. 21). As the gate length is decreased, total generation and, therefore, leakage current is reduced. This is due to the reduced volume, since the total generation is obtained by volume integration. Here we consider only the bulk recombination, and not the surface one. For 0.15 μm device I_{on}/I_{off} separation improves to four decades. Subthreshold slope (SS) varies from 150mV/decade for the shortest

Lg to 190 mV/decade for the longest. Note that these values of SS are idealized because we have assumed that the oxide-InSb interface is perfect, i.e. it does not contain any traps, unlike the experimental device in [6]. For the shortest device SS is larger than in comparable Si devices in large part due to the thicker gate oxide. Similar behavior was observed in quantum-well devices [5]

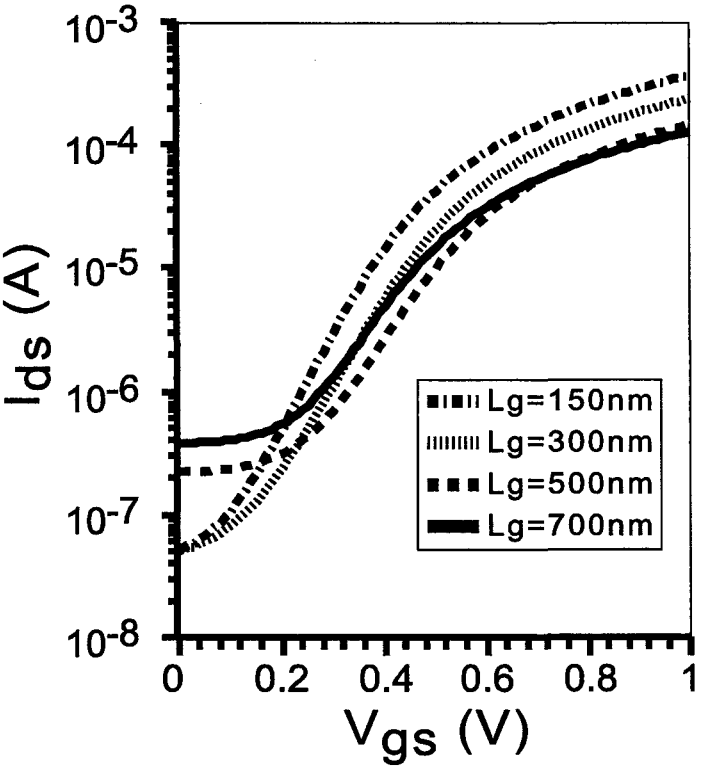

Figure 21: I_{on}/I_{off} vs. gate length show more than 3 decades in I_{on}/I_{off} separation, which is satisfactory performance; I_{off} is I_{ds} at zero volts, I_{on} at 1V.

High electron saturation velocity in InSb makes unity current gain frequency extremely high. Fig. 22 illustrates this point. Unity gain frequency is calculated from the equation:

$$f_T = \frac{g_m}{2 \cdot \pi \cdot C} \tag{33}$$

Where g_m is transconductance and C is input capacitance. They are calculated through ac simulation in DESSIS.

Gate voltage is swept up to 3V, while V_{ds} is held at 0.5V to maximize transconductance. The largest value of f_T for 0.7 μm device is about 100 GHz, attained at V_{gs} of 1.2V. This compares well with the simple calculation assuming velocity saturation throughout the channel:

$$f_t = \frac{v_{sat}}{2\pi L_g} = \frac{5 \cdot 10^5}{2\pi \cdot 0.7 \cdot 10^{-6}} = 113GHz \tag{34}$$

Also, this number is close to the number reported by Ashley et al [6]. Their simulated 0.7um device attained about 80 GHz max f_T .Experimental results show fall of in the unity gain frequency with the gate voltage, while our data levels off after the peak. This is due to the mobility modeling due the vertical field dependence. As mentioned earlier, reliable measured data is not available. However, for our conclusions, we are mostly interested in the peak value, which is aligned with the simulations reported in [6].

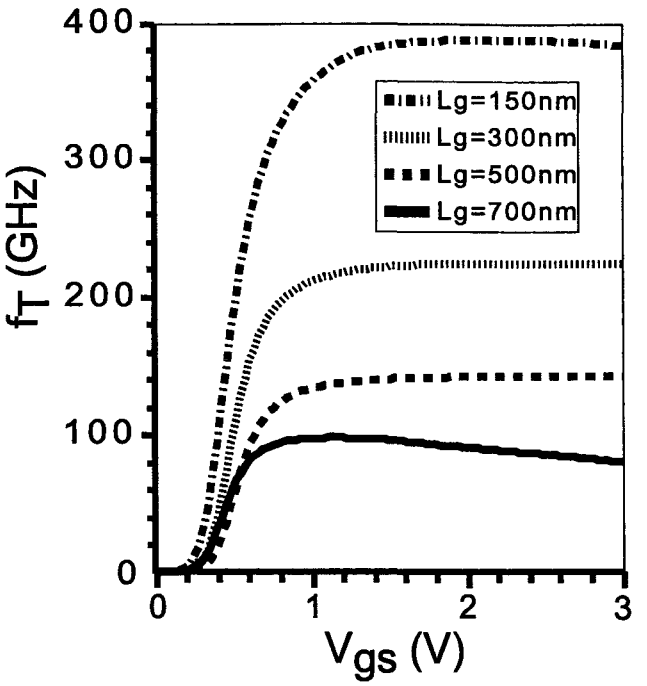

Figure 22: f_T vs. gate length and gate voltage V_{gs}. inSb shows very high values for the unity gain frequency.

For 0.15 μm gate length f_T reaches 387 GHz. In the Fig. 23 InSb MOSFET values of f_T are compared with Si and InP devices. As can be readily observed, values of f_T for InSb MOSFET are much larger than the ones for silicon MOSFET of comparable gate length. InSb scales approximately as $75*L_g^{-0.88}$, which is a little bit less than what is expected from velocity saturation limit. We can see this by observing that InSb curve is slightly below saturation velocity limit curve in the Fig 23. If the same trend is extrapolated, f_T-s around 700 GHz are expected for L_g=75 nm and 1.4 THz for Lg=45 nm. These numbers compare well with the predictions in [6]. It should be

emphasized that these are only rough estimates since quantum effects and quasi-ballistic effects have not been included.

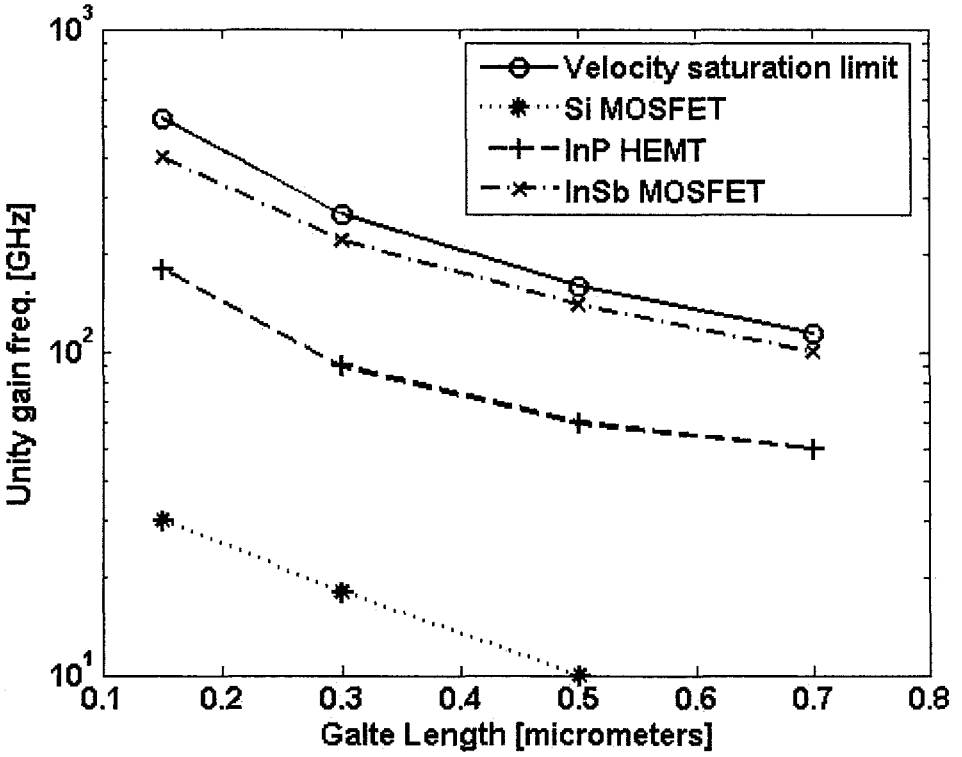

Figure 23: Extrapolated f_T demonstrates InSb is the most promising material for high speed applications. Numbers for Si and InP from [5]

8. Quantum mechanical effects

As we have mentioned earlier, all results presented so far were obtained using DESSIS software. It is a very powerful and accurate simulator for silicon devices, but it allows limited user modifications. It was not possible to modify it to account for nonparabolic nature of the Schrödinger equation for InSb. Hence, results obtained in this section (Section 8) are based on nanoMOS simulator [19].

Effects of the quantization on the device performance have been investigated in literature [39, 40]. However, our goal is to analyze impact of non-parabolicity on the quantum mechanical effects in InSb which was not done self-consistently [7].

8.1. Non-parabolic quantum mechanical model

For parabolic materials, effective mass used in transport and in Schrödinger equations is constant and calculated at the bottom of the conduction band, typically at the bottom of the Γvalley. However, for non-parabolic materials effective mass will be function of both position and energy and is given by [41]:

$$\frac{1}{m(E,r)} = \frac{P^2}{\hbar^2}\left[\frac{2}{E+E_g-V(r)} + \frac{1}{E+E_g+\Delta-V(r)}\right] \tag{35}$$

Where P is Kane momentum matrix element, V is confinement potential, E_g is bandgap and delta is split-off band energy. Since the effective mass will be varying within the well, Ben Daniel-Duke form of the Hamiltonian ought to be used [41]:

$$H = -\frac{\hbar^2}{2}\nabla_r\left(\frac{1}{m(E,r)}\right)\nabla_r + V(r) \tag{36}$$

For 1D case and using chain rule we obtain:

$$-\frac{\hbar^2}{2}\frac{d}{dx}\frac{1}{m}\frac{d}{dx}\psi - \frac{\hbar^2}{2m}\frac{d^2}{dx^2}\psi + V\psi = E\psi$$
$$\frac{\hbar^2}{2m^2}\frac{d}{dx}m\frac{d}{dx}\psi - \frac{\hbar^2}{2m}\frac{d^2}{dx^2}\psi + V\psi = E\psi \tag{37}$$

This Hamiltonian is implemented in nanoMOS using center-difference equations to approximate first and second derivatives. Approximation of eq. (37) by center differences gives [42]:

$$-\frac{\hbar^2}{2m_i(\partial y)^2 q}\left[\begin{array}{l}\left(1+\frac{m_{i+1}-m_{i-1}}{4m_i}\right)\cdot\psi_{i-1}- \\ 2\cdot\psi_i+\left(1-\frac{m_{i+1}-m_{i-1}}{4m_i}\right)\cdot\psi_{i+1}\end{array}\right]+V\cdot\psi_i=E\cdot\psi_i \qquad (38)$$

Where dy is mesh size in the direction perpendicular to the oxide-InSb interface, ψ is a wavefunction evaluated at different nodes and E is energy eigenvalue. We assume boundary condition for the wavefunction to be zero inside the oxide, i.e. no penetration.

Eq. (38) is implemented in nanoMOS and solved iteratively with the eq. (35). Initially, we assume energy E in eq. (35) to be zero and calculate mass at each mesh point. These values of the mass are inserted in the eq. (38) from which we calculate energy eigenvalue. If the error between calculated energy and the value from the previous pass is greater than 1%, loop is repeated and new effective masses are calculated from the eq. (32) with new energy E. Once E and m converge to their final values, average effective mass for the whole channel is calculated and used for effective density of states. Since variation is small in either x or y direction, average for a whole device is calculated. Procedure is justified as long as the variation of the effective mass along the channel (in x-direction) is small, which is the case as will be shown later. Finally, it should be noted that nanoMOS solves Poisson's equation

simultaneously with the above quantum mechanical equations so that a simultaneous solution for electron concentration and energy levels is obtained.

8.2 Impact of non-parabolicity on energy levels in finite rectangular quantum wells

In order to asses impact of nonparabolicity in the Schrödinger equation, we compared solutions for finite rectangular quantum wells with parabolic and non-parabolic E-k relationship. This example is chosen since it is the only one that can be analytically solved with non-parabolic mass. Energy eigenvalues are solutions of the transcendental equation:

$$k_2\sqrt{\frac{m_b}{m_w}}\frac{w}{2}=\frac{m_b k_1}{m_w}\frac{w}{2}\tan\left(k_1\frac{w}{2}\right) \qquad (39)$$

Where m_b and m_w are masses in the barrier and in the well respectively, while

$$k_1=\sqrt{\frac{2m_w E}{\hbar^2}}$$

$$k_2=\sqrt{\frac{2m_b(V_0-E)}{\hbar^2}}$$

are wavevectors in the well and the barrier. W is well width and V_0 is well depth. First solution (smallest energy E) of the eq. (39) represents ground state. For the parabolic E-k case, two masses are identical and equal to 0.014 m_e. For non-parabolic case, we substitute masses with the eq. (35) using appropriate potential.

Ground state energies for different well widths are plotted in the Fig. 24 below. Well depth is 1eV.

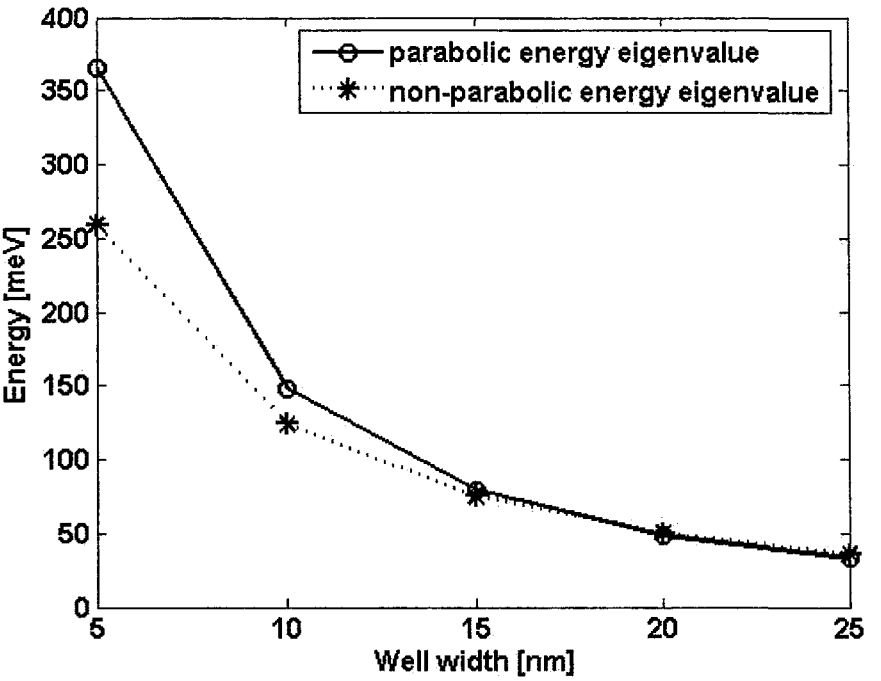

Figure 24: 1eV rectangular quantum well ground states; below 15nm nonparabolicity has to be taken into account.

For well widths smaller than 15 nm, impact of non-parabolicity on the ground state becomes appreciable. Band non-parabolicity reduces energy, and in the case of 5 nm well, it can lower it by almost 30%. Given this, appropriate quantum mechanical modeling needs to be implemented in nanoMOS Matlab code.

To better guide our intuition, the expected effect of the non-parabolicity is shown in Fig. 25. E-k diagram on top represents parabolic E-k relationship and

the one on the bottom non-parabolic. Total energy of the system in the quantum well is given by:

$$E = \frac{\hbar^2 k^2}{2m(E)} + E_n \qquad (40)$$

Where E_n is the energy eigenvalue, i.e. solution to the Schrödinger equation, and m is energy dependent effective mass. Since non-parabolicity increases effective mass, total energy will be lower for a given k, compared to the parabolic case, as shown in the Fig. 25 Also, because of increased effective mass density of states is increased, which can be interpreted as reducing apparent (effective) band gap. We would, therefore, expect non-parabolicity to increase current and reduce threshold voltage, as reported in the literature [7].

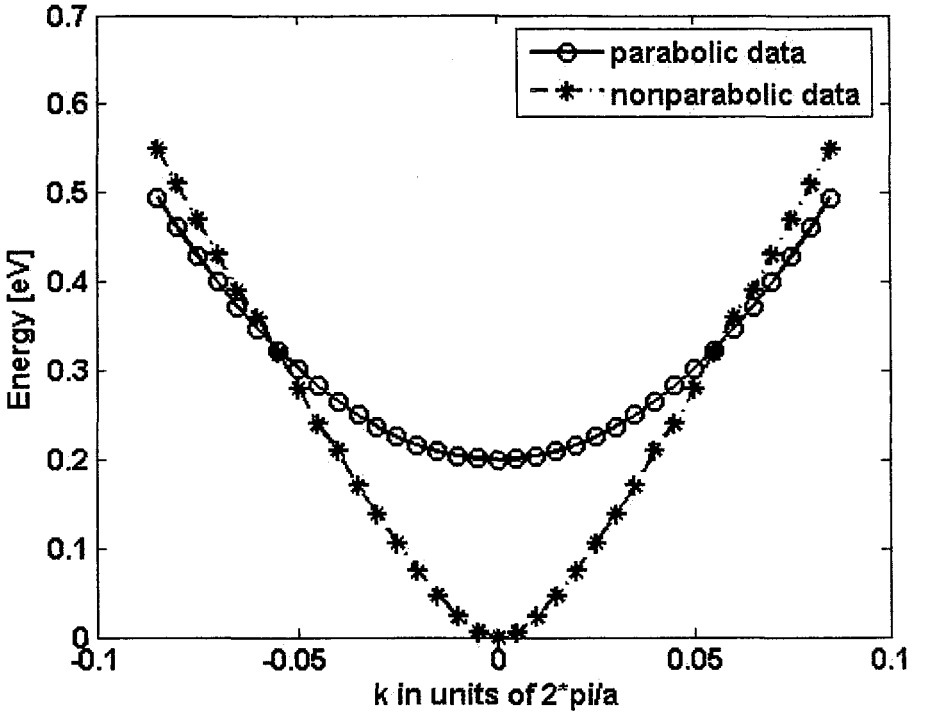

Figure 25: Comparison of parabolic and nonparabolic E-k diagram. Assumption of parabolicity results in large errors in energy

8.3 Dual gate InSb MOSFET device

One of the trends in MOSFET design is to use multiple gates [43], so that there are single-gate, dual-gate, tri-gate and all-around gate designs. In general, this is supposed to provide larger current from the same device layout by having several inversion layers in parallel. These devices are also supposed to have better gate control of the channel charge which should increase transconductance. It is expected that deeply scaled devices will utilize one of these gate designs. NanoMOS is set up to analyze dual gate

structures so we can easily investigate how the introduction of InSb will affect dual-gate MOSFET performance.

Cross-section of the device under investigation is shown in Fig. 26. Gate length is 40 nm, and source and drain regions are 20 nm long. At this length drift-diffusion assumptions are still reasonably accurate [44], i.e. we can neglect ballistic transport. For insulator we use hafnium oxide, 3nm thick with relative dielectric constant of 24. We examine three different body thicknesses, 5, 10, 12 and 15 nm. Based on silicon FinFETs, generally it is expected that the body thickness should be no bigger than half the gate length, which fits numbers used here [45].

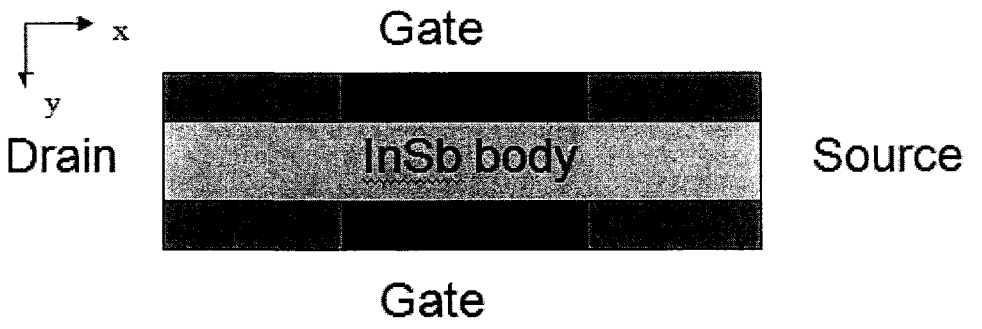

Figure 26: Dual gate InSb MOSFET cross section

8.4 Device results

Following the procedure given in [44] gate work function is adjusted to produce 0.1 A/m I_{off} at zero volts gate bias . V_{ds} is held at 0.5V. InSb body is kept intrinsic, while source and drain regions are doped to 10^{19} cm^{-3}. Two different cases are simulated: parabolic, with constant effective mass at the bottom of the Γvalley (0.014) and non-parabolic case where effective mass is a function of energy, as explained earlier. Plot of the I_{ds} vs. V_{gs} is shown in Fig. 27, V_g is swept in 0.05V steps up to 1.05V. As expected, including non-parabolicity in simulation reduces threshold voltage dramatically, by 0.6V.

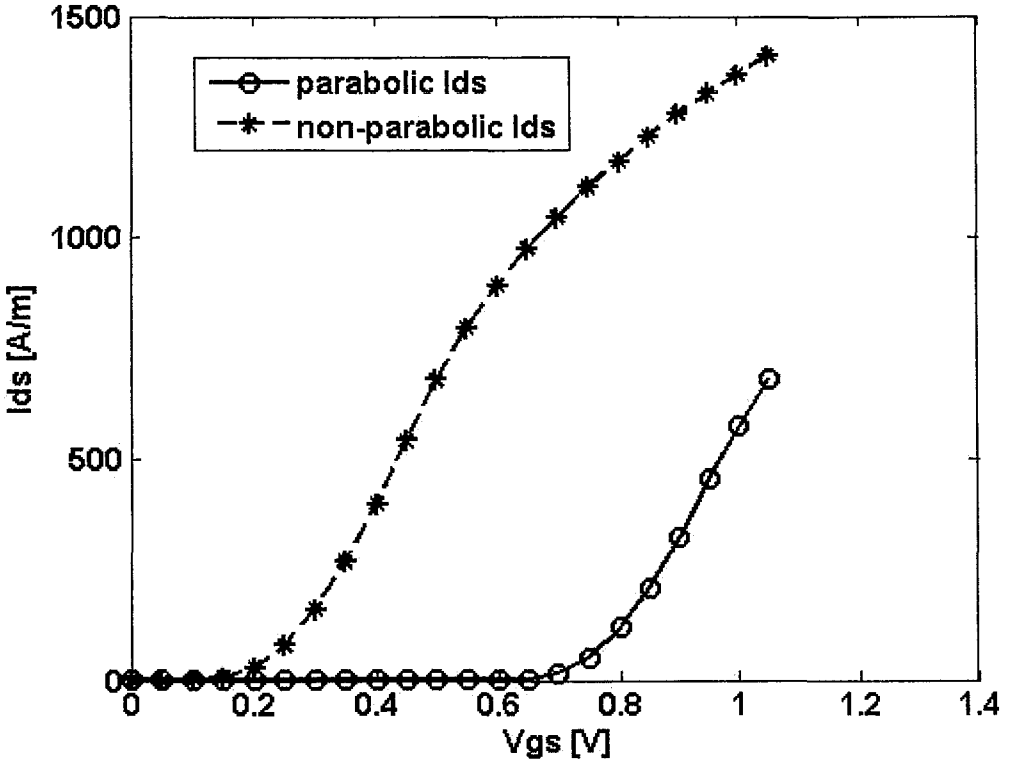

Figure 27: I_{ds} vs. V_{gs} for V_{ds}=0.5V show large shift in the threshold voltage due to the nonparabolicity.

Same data as in the Fig. 27 can be plotted on the logarithmic scale as shown in Fig. 28. From Fig. 28 subthreshold slope SS is calculated and it is roughly the same for both parabolic and nonparabolic cases, i.e. 80mV per decade for parabolic and 90 mV per decade for nonparabolic case. Since the standard Si MOSFET has a lower limit of about 2.3kT/q= 60 mV per decade our results are acceptable.

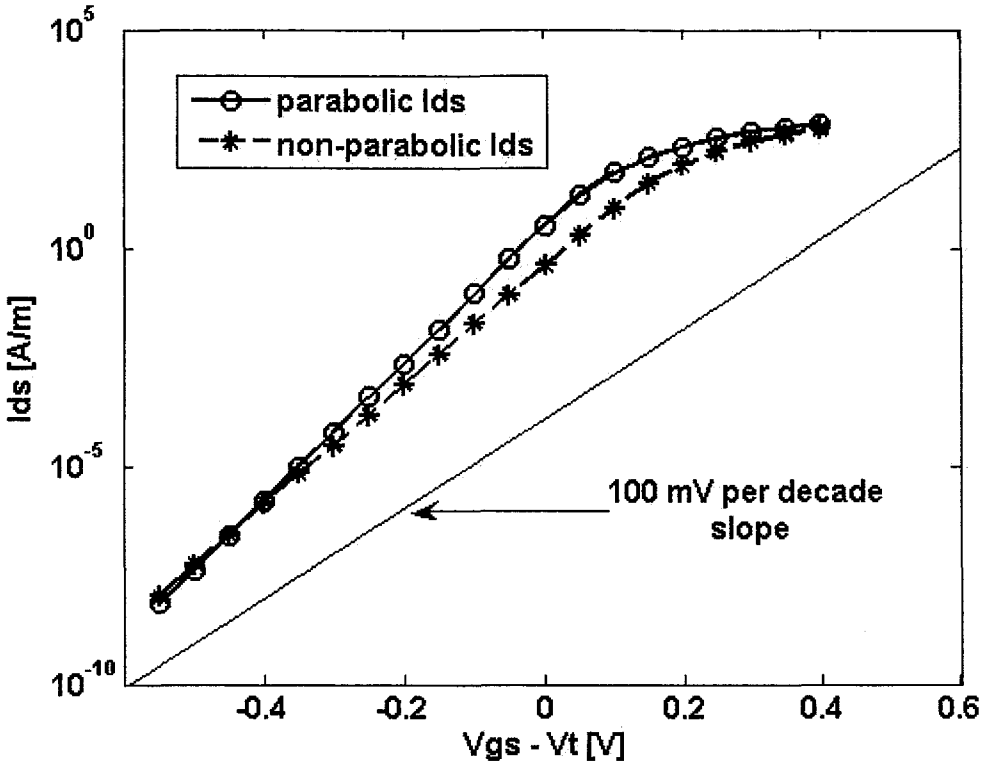

Figure 28: Subthreshold I-V curves and subthreshold slope calculation; nonparabolicity slightly increases subthreshold slope.

As noted earlier, we neglect variations of the effective mass along the channel in the calculation of the density of states. Since InSb body is thin and intrinsic, variation in potential in the perpendicular direction is small as well, resulting in the small change in the effective mass. Fig. 29 shows effective mass variation from source to drain contact at V_{gs}=1.05V and V_{ds}=0.5V. Effective mass variation is much less than one percent, so procedure outlined in the earlier section is justified. Note that although variation is minimal,

effective mass is around 2.5 times larger than at the bottom of the Γvalley where m= 0.014.

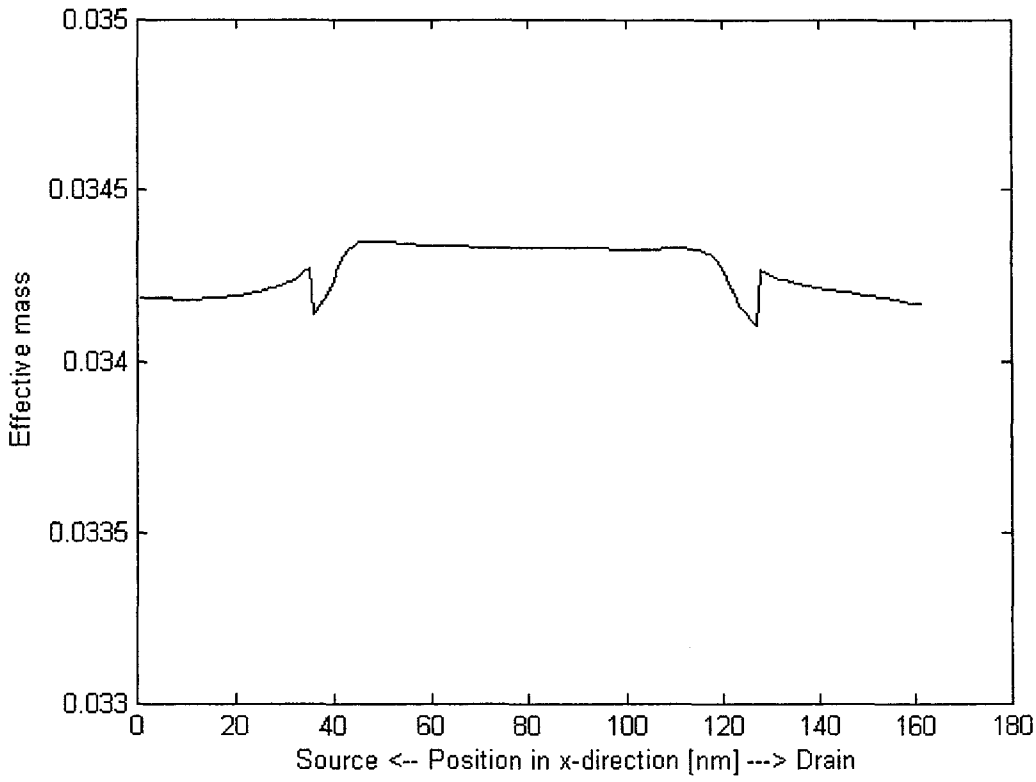

Figure 29: Effective electron mass along the channel (in units of free electron mass) is virtually constant, though different from the bottom of the Γvalley value.

Vertical cross section of the potential in the centre of the device is in the Fig. 30. Body thickness is 5 nm and bias applied is V_{gs} = 1V and V_{ds} =0.5V. We can see that variation is less than 10%, which explains small variation in the effective mass. In principle, for thin body devices, due to small variation in the effective mass, ordinary form of the Schrödinger equation could have been

used, rather than Ben Daniel-Duke. However, iterative procedure to calculate mass and energy eigenvalue proposed in section 8.1. should still be followed.

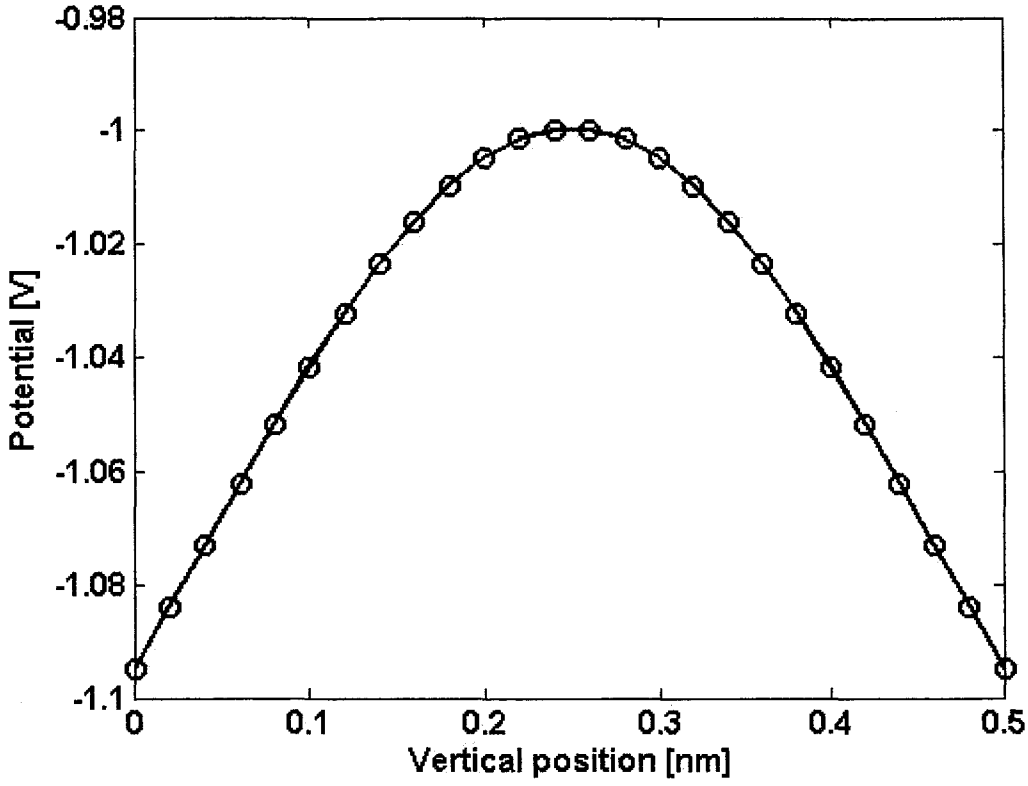

Figure 30: Variation of the potential in InSb dual gate MOSFET in the vertical direction is within 10%.

Fig. 31 shows change in peak transconductance g_m, as a function of body thickness. Somewhat surprisingly, g_m increases as the body thickness increases. Reduction in transconductance for thin body dual gate FET results from mixing of the inversion layers from two gates in Fig. 26, similar to what

was observed in very narrow silicon Fin-FETs [25]. This effect is illustrated in Fig. 32 , which represents variation of electron concentration along the vertical cross section of the device. Channels of the individual gates are merged into one, resulting in the peak of the concentration in the center of the device, rather than at the InSb-insulator interface. Fig. 32 also compares parabolic and nonparabolic behavior. Gate work function is chosen to achieve similar overdrive for both parabolic and nonparabolic cases.

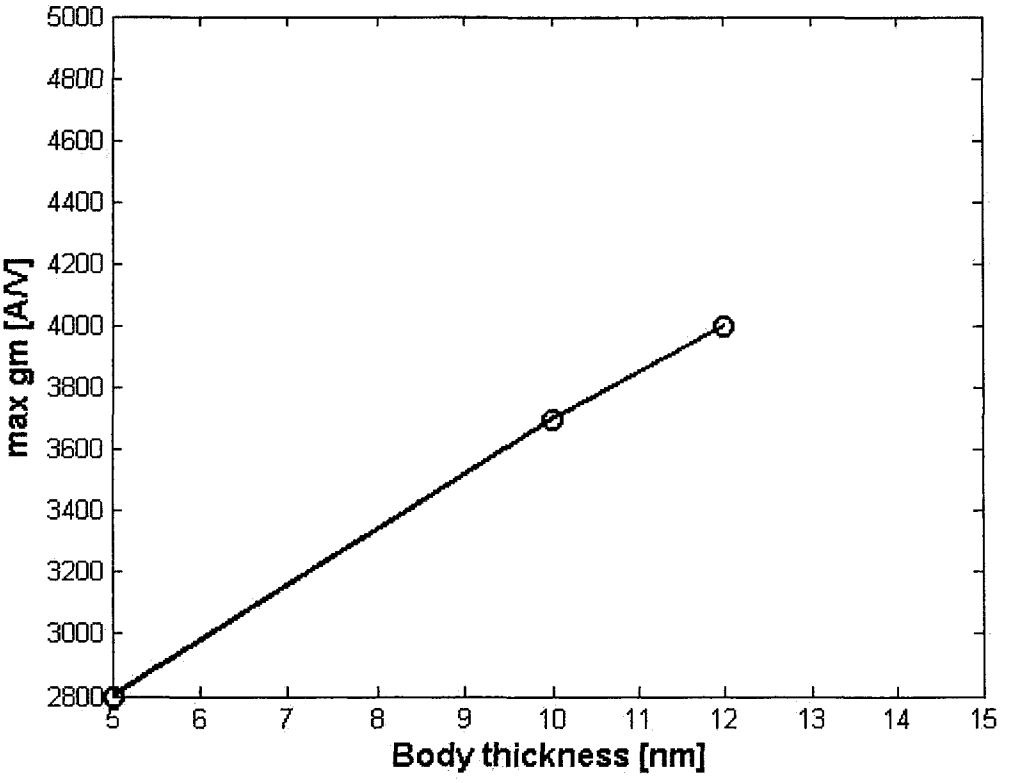

Figure 31: Max. Transconductance vs. body thickness shows substantial decrease in the transconductance with the reduced body thickness due to the mixing of the inversion layers.

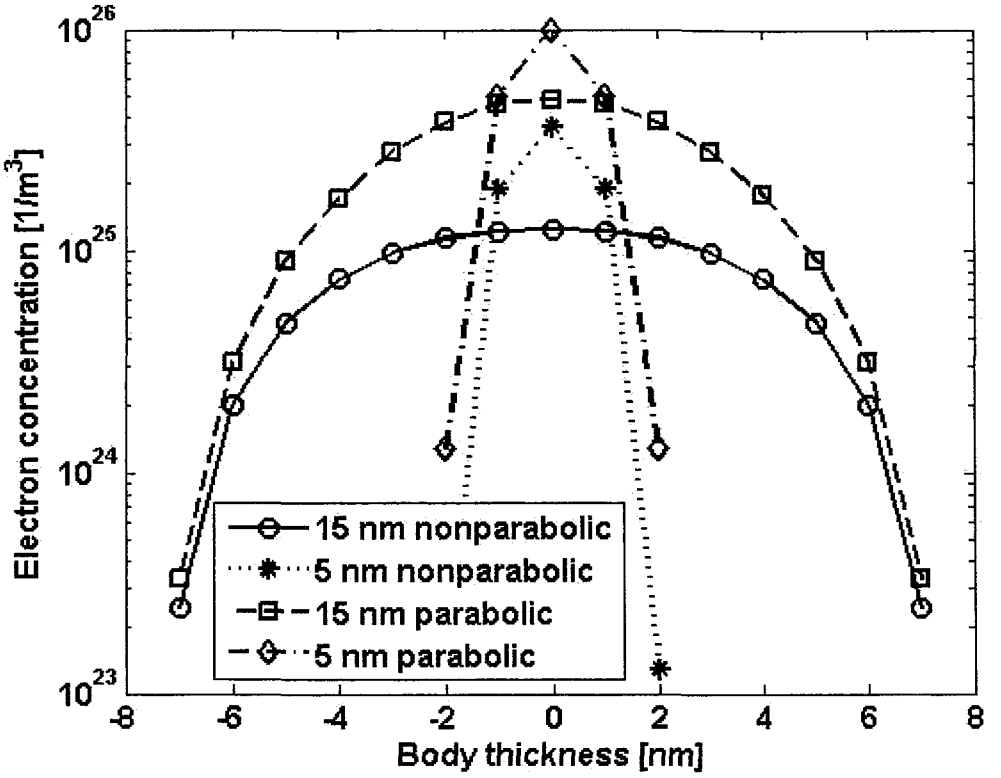

Figure 32: Electron concentration vs. vertical position for 5 and 15 nm thick device; maximum concentration occurs in the middle of the device due to the QM effects.

As can be seen from Fig. 32, even for a relatively thick body (15 nm) two channels are merged. This is actually expected behavior, since body is undoped. This effect is not related to the nonparabolicity, but it is rather consequence of the quantum mechanical behavior [25]. Classical treatment would yield maximum carrier concentration immediately at the surface of the oxide, no matter how thin we make this device. Due to the nature of the wave function inside quantum well, peak electron concentration will be shifted some

distance from the surface. Hence, if body is thin enough, two peaks will merge as shown in the Fig. 31. Also from Fig. 32 we see that nonparabolicity has no effect on the shape of the electron concentration curve. We could not simulate thicker body due to convergence problems.

Reduction in the transconductance, in turn, will decrease unity gain frequency. Plot of f_T vs. body thickness is given in the Fig. 33 for 40 nm gate length. Maximum f_T for 5 nm thick body device is about 630 GHz, for 10nm body it is 790 GHz and 12 nm about 760 GHz indicating that an optimum value may have been reached. However, as the body thickness is increased, it was more difficult for the simulations to converge since the tightening of the convergence criteria was required in order to avoid non-physical results Devices with body thickness larger than 13 nm showed some non-physical behavior such as increase in f_T at some V_{gs} value after the first peak value was reached; normally, we expect to see a continued fall-off or saturation of g_m value. These could be removed with further tightening of the convergence criteria, however, due to the lack of the computing memory this was not possible.

But our main interest was the value of the peak, which we believe was simulated reliably. Simple formula for unity gain frequency, assuming velocity saturation, would give 1.9 THz for 40 nm device:

$$f_t = \frac{v_{sat}}{2\pi L_g} = \frac{5 \cdot 10^5}{2\pi \cdot 40 \cdot 10^{-9}} = 1.98 THz$$

It can be concluded that mixing of the inversion layers substantially reduces saturation velocity limited f_T by about 50%

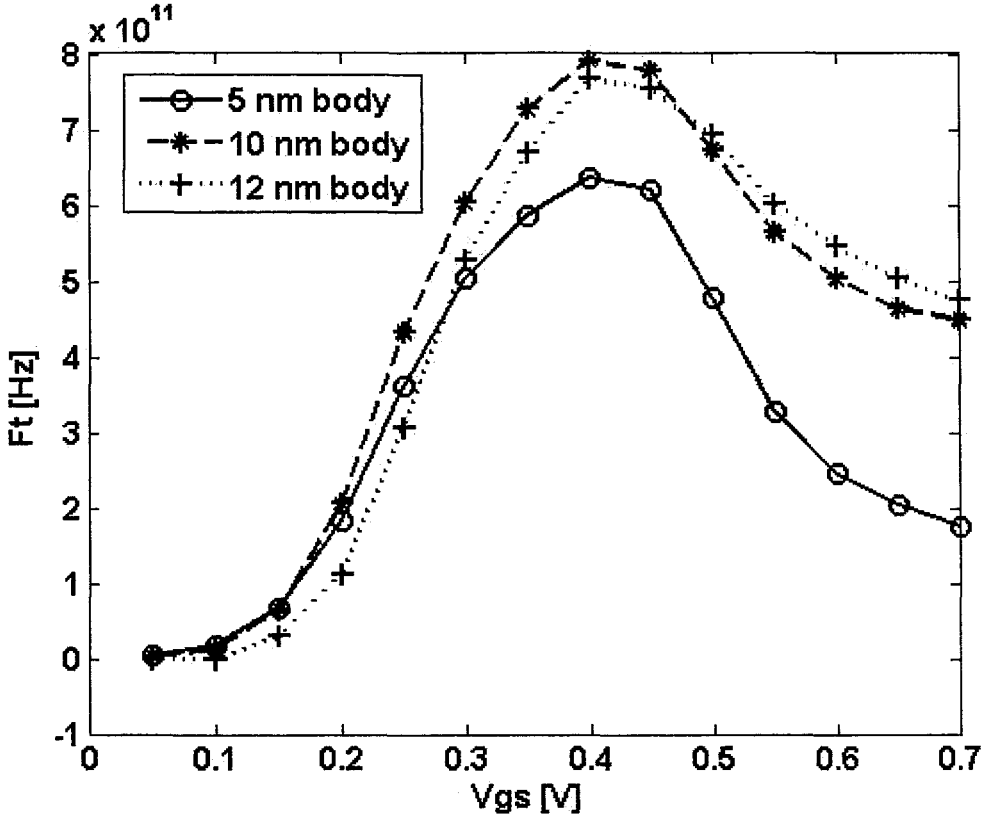

Figure 33: Unity gain frequency is directly proportional to the body thickness due to the mixing of the inversion layers.

9. Conclusions

Objective of this project was to develop critical InSb models that can then be used in designing and modeling devices, diodes and MOSFETs and to demonstrate feasibility and advantages of InSb MOSFET device over comparable silicon technology, through different metrics, such as I_{on}/I_{off} ratio and unity gain frequency.

The most challenging part of the project, and likewise, the most important contribution, is in the development of the appropriate and self-consistent physical and mathematical models for InSb. Since previously most important application were photo-diodes, InSb was partially characterized at 77K, but room temperature data, necessary to model practical transistor devices, is lacking. We have shown that very high nonparabolicity and degeneracy affect different physical properties, such as electron concentration, Auger generation and quantum mechanical behavior. We have seen that current state of the art treatment of non-parabolicity is either constrained to the low nonparabolicity or it does not treat it in a self-consistent manner, as needed in quantum mechanics.

Another difficulty is the numerical stability of the models and implementation. Currently available commercial simulators, such as DESSIS, are not designed for highly non-parabolic materials. To account for the higher order terms in the power series expansion of the electron concentration (eq

(13)), we had to artificially modify effective mass through the pmi routine. Auger mechanism is modeled through the exponential function, which worked fine for the diodes, but MOSFET simulations became numerically unstable. Hence, we had to come up with an alternative model for Auger generation-recombination. In the nanoMOS tool, we had to develop iterative simulation procedure for the calculation of the energy eigenvalue and the effective mass in the self-consistent manner. So instead of solving coupled Poisson and Schrodinger equation, like it is done for the parabolic materials, tool solves coupled effective mass, Schrodinger and Poisson equation. Additional layer of complexity causes numerical instability for the thicker device body and higher doping.

After these problems were surmounted, we demonstrated that InSb MOSFET devices can successfully be simulated in the commercially available tools. Results, like differential resistance, compare well with the measurements. Others, like unity gain frequency, compare well with other theoretical calculations.

Exclusion/extraction principle is shown to dramatically improve device performance, reducing off state current by more than a decade. We were able to reduce the leakage current so the ratio of the on to off current was about three decades, minimum requirement for the digital design. However, short channel effects become important already at gate lengths of around 300nm, so we also demonstrate that appropriate scaling rules can be developed. By applying those rules, our 150nm gate length device was able to reach unity

gain frequency of almost 400 GHz. This number is much higher than silicon for the comparable technology and is also higher than other III-V compounds, like InP. Appropriate methodology for nonparabolic quantum mechanical treatment is developed and implemented in nanoMOS. We demonstrated, through simulation, operational dual-gate InSb FET device. We show that nonparabolicity has a large effect on quantum mechanical behavior, reducing energy eigenvalues relative to parabolic results, and increasing effective mass. In terms of the I-V characteristics, nonparabolicity will reduce threshold voltage substantially and has to be taken into account in the Schrodinger equation.

We also identified major roadblocks for the further development and scaling of the dual gate InSb MOSFET, like mixing of the inversion layers and the reduction in the transconductance and the unity gain frequency.

This project was initiated to explore potential of InSb as a semiconductor material for MOSFET technology. We have demonstrated that suitable devices operating at the room temperature can be designed with the performance superior to silicon, using unity gain frequency as a figure of merit. However, downside of these devices is the lack of the practical p-type device, rendering CMOS technology unusable. Also, requirement for the additional power supply to reversely bias body would increase the cost. Given these, we believe that although InSb devices show promising performance, they may not be suitable for mass production of the cheap computing devices, such as PCs. In fact, we see the future for InSb in the specialized applications where the performance is

of the essence, and the cost concerns are secondary. In these niche applications, InSb MOSFET would perform well.

Last, but not the least, current infrastructure in the industry is geared toward silicon, making any new technology change very expensive. However, this problem is common to all new materials, be it InSb, carbon nanotubes or other III-V compounds.

As for the future investigations, this project can be continued by devising suitable design rules for a double gate InSb MOSFET to remedy negative effects of the mixing of the inversion layers and resulting reduction in the transconductance and the unity gain frequency. These may include appropriate thickness of the body and/or channel doping. More efficient and robust computational engine may be integrated in Matlab to enhance convergence. Also, further scaling of the device below 40 nm would require ballistic transport, rather than drift-diffusion, which was main tool for this project. Ballistic transport equations would need to be modified to account for the nonparabolicity of the material. Quantization in the lateral direction may also be necessary. Also, possibility of designing a p-type device through e.g. strain engineering can be explored.

References

[1] Eiichi Sano and Taiichi Otsuji , "HEMT-Based nanometer devices toward terahertz era", International Journal of High Speed Electronics and Systems, Vol. 17, No 3, 2007, pp 509-520

[2] Y. Nishi, R. Doering "Handbook of semiconductor manufacturing technology", Marcel Dekker, 2000, p 291

[3] http://download.intel.com/research/silicon/Gordon_Moore_ISSCC_021003.pdf

[4] http://www.itrs.net/Links/2008ITRS/Update/2008_Update.pdf

[5] Robert Chau et al, "Benchmarking Nanotechnology for High-Performance and Low-Power Logic Transistor Applications", IEEE Transactions on Nanotechnology, Vol. 4, 2005, pp 153-158

[6] T. Ashley, et al." High-Speed, Low-Power InSb Transistors," IEDM Technical Digest, 1997, pp. 751-754

[7] A. Khayer et al "Performance of a n-Type InSb and InAs Nanowire Field Effect Transistor", IEEE Transactions on Electron Devices, 2008, pp 2939-2945

[8] E. Sijercic,; B. Pejcinovic, "Investigation of scaling of InSb MOSFETs through drift-diffusion simulation", Solid-State Electronics, Vol.50, 2006, pp 1634-1639

[9] A. Akturk , "Terahertz current oscillations in single-walled zigzag carbon nanotubes", Physical Review Letters, Vol. 98, 2007, pp 166803-1 to 166803-4

[10] H. W. Postma, et al. "Carbon Nanotube Single-Electron Transistors at Room Temperature", Science, 293, 2001, pp 76-79

[11] D. H. Kim and J. A. del Alamo "Beyond CMOS: Logic Suitability of $In_{0.7}Ga_{0.3}As$ HEMT" CS MANTECH Conference,. April 24-27, 2006, Vancouver, British Columbia, Canada, pp 251-254

[12] Minjoo L. Lee and Eugene A. Fitzgerald "Hole mobility enhancements in nanometer-scale strained-silicon heterostructures grown on Ge-rich relaxed $Si_{1-x}Ge_x$", Journal of Applied Physics, Vol. 94, 2003, pp 2590-2596

[13] S. Tiwari, "Hole mobility improvement in silicon-on-insulator and bulk silicon transistors using local strain" International Electron Devices Meeting. IEDM Technical Digest, 1997, p. 939

[14] David Chinnery and Kurt Keutzer, "Closing the gap between ASIC & Custom: tools and techniques for high-performance ASIC design", Kluwer Academic Publishers, 2003, p. 13

[15] E. Sijercic, K. Mueller, and B. Pejcinovic, "Simulation of InSb devices using drift-diffusion equations," Solid-State Electronics, Vol. 49, 2005, pp. 1414-1421

[16] E. Sijercic, K. Mueller, and B. Pejcinovic " Drift-diffusion Simulation of InSb devices", MELECON 2004, Vol. 1, pp 43-46

[17] E. Sijercic and B. Pejcinovic "Simulation of non-parabolicity effects in dual-gate InSb MOSFETs", to be submitted for publication in Solid State Electronics

[18] Dessis device simulation software, release 7.5. ISE Integrated Systems Engineering AG, Zurich, Switzerland, 2001.

[19] Zhibin Ren; Sebastien Goasguen; Akira Matsudaira; Shaikh S. Ahmed; Kurtis Cantley; Mark Lundstrom; Xufeng Wang NanoMOS; DOI: 10254/nanohub-r1305.7. 2006

[20] M. Levinshtein et al., editors, "Handbook Series on Semiconductor Parameters", London: World Scientific, Vol. 1, 1996, pp. 191-213

[21] Van E. Wood "Evaluation of some transport integrals.IV.Nonparabolic bands", Journal of Applied Physics, Vol. 44, no. 4, 1973, pp 1515-1517

[22] A.W. Smith and K.F. Brennan "Non-parabolic hydrodynamic formulations for the simulation of inhomogeneous semiconductor devices", Solid state electronics, Vol. 39, 1996, pp 1659-1668

[23] Mark Lundstrom and Jing Guo, "Nanoscale transistors; device physics, modeling and simulation", Springer, 2006

[24] Xiao Zhixiong, et al, "Calculation and Analysis of the Intrinsic Carrier Concentration and the Einstein Relation for Heavily Doped Silicon from 77 K to 300 K", Japanese journal of applied Physics, Vol. 35, 1996, pp 1599-1604

[25] M. Poljak et al "Quantum Confinement and Scaling Effects in Ultra-Thin Body Double-Gate FinFETs", MIPRO 2009

[26] T. Ashley et al. "A heterojunction minority carrier barrier for InSb Devices". Semicond Sci Technol 1993;8:S386–9

[27] J.C. Cao and X.L. Lei "Nonparabolic multivalley balance equation approach to impact ionization: Application to wurtzite GaN", The European Physical Journal B, B7, 1999, pp 79-83

[28] X. F. Wang and X.L. Lei "Impact ionization in balance equation theory", J. Phys: Condens Matter, Vol. 7, 1995, pp 7871-7878

[29] A. R. Beattie , A. M. White "An analytic approximation with a wide range of applicability for electron initiated Auger transitions in narrow-gap semiconductors". Journal of Applied Physics Vol. 79, 1996; pp802–13.

[30] T. Ashley , C.T. Elliott "Operation and properties of narrow-gap semiconductor devices near room temperatures using non-equilibrium techniques". Semicond Sci Technol Vol. 6, 1991;pp C99–C105.

[31] G. J. Nott et al, "Direct determination of Shockley-Read-Hall trap density in InSb/InAlsSb detectors", J. Phys.: Condens. Matter, Vol. 12., 2000, pp. L731-L734

[32] G.A.M. Hurkx, D.B.M. Klaasen, M.P.G. Knuvers, "A New Recombination Model for Device Simulation Including Tunneling," IEEE Trans. Electron Devices, Vol. 39, no. 2, 1992, pp. 331-338

[33] E.O Kane, "Theory of Tunneling," Journal of Applied Physics, Vol. 32, no. 1,1961, pp. 83-91

[34] W. Zawadski, Advances in Physics, Vol. 23, 3/ 4, 1974.

[35] Arora et al, "Electron and hole mobilities in Silicon as a function of concentration and temperature", IEEE Trans. Electron Devices, Vol. ED-29, 1982, pp. 292-295

[36] C. Canali et al "Electron and hole drify velocity measurements in Silicon and their empirical relation to the electric field and temperature", IEEE Trans. Electron Devices, Vol. ED-22, 1975 pp. 1045-1047

[37] T. Ashley, et al. "Ambient temperature diodes and field-effect transistors in InSb/$In_{1-x}Al_xSb$," Applied Physics Letters, Vol. 59, no. 14, 1991 pp. 1761-1763

[38] Yuan Taur and Tak Ning "Fundamentals of modern VLSI devices" Cambridge University Press, 2001

[39] K.D. Cantley et al. "Performance Analysis of III-V Materials in a Double-Gate nano-MOSFET". IEDM 2007. IEEE International 2007: pp 113-116

[40] L. Samuelson, "Semiconductor nanowires as an approach towards electronic and photonic devices", Indium Phosphide and Related Materials, IPRM 2008. 20th International Conference on, pp 1

[41] Y. Li et al, "Energy and coordinate dependent effective mass and confined electron states in quantum dots", Solid State Communications Vol. 120, 2001; pp 79-83

[42] D. Eberly "Derivative approximation by Finite Differences" www. geometrictools.com

[43] Maesoon Im, "Multiple-Gate CMOS Thin-Film Transistor With Polysilicon Nanowire", IEEE Electron Device Letters 2009, pp 102-105

[44] J. Wang and M. Lundstrom, "Ballistic Transport in High Electron Mobility Transistors", IEEE Transactions on Electron Devices Vol. 50, 2003, pp 1604-1609

[45] Wenwei Yang, "Scaling Theory for FinFETs Based on 3-D Effects Investigation", IEEE Transactions on electron devices, Vol. 54. 2007, pp 1140-1147

[46] Jean-Luc Thobel et al, "Monte Carlo simulation of electron transport in narrow gap heterostructures" Journal of applied physics, Vol. 92, 2002, pp 5286 – 5295

[47] D. C. Herbert et al "Monte Carlo Simulations of High-Speed InSb–InAlSb FETs", IEEE Transactions on Electron Devices, Vol. 52, 2005, pp 1072-1078

[48] Zhibin Ren, Ramesh Venugopal, Sebastien Goasguen, Supriyo Datta, and Mark S. Lundstrom "nanoMOS 2.5: A Two -Dimensional Simulator for Quantum Transport in Double-Gate MOSFETs," IEEE Transactions on

Electron Devices, special issue on Nanoelectronics, Vol. 50, 2003, pp. 1914-1925