Title
Reinventing the PN Junction: Dimensionality Effects on Tunneling Switches

Author
Agarwal, Sapan

Publication Date
2012

Peer reviewed|Thesis/dissertation

Reinventing the PN Junction: Dimensionality Effects on Tunneling Switches

By

Sapan Agarwal

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California Berkeley

Committee in charge:

Professor Eli Yablonovitch, Chair
Professor Sayeef Salahuddin
Professor Junqiao Wu

Spring 2012

Reinventing the PN Junction: Dimensionality Effects on Tunneling Switches

Abstract

Reinventing the PN junction: Dimensionality Effects on Tunneling Switches

by

Sapan Agarwal

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Eli Yablonovitch, Chair

Tunneling based field effect transistors (TFETs) have the potential for very sharp On/Off transitions. This can drastically reduce the power consumption of modern electronics. They can operate by either electrostatically controlling the thickness of the tunneling barrier or by exploiting a sharp step in the density of states for switching. We show that current TFETs rely on controlling the thickness of the tunneling barrier but they do not achieve the desired performance. In order to get better performance we need to also exploit a sharp step in the density of states.

In order to have a sharp density of states turn on, a variety of non-idealities need to be accounted for. A number of effects such as thermal vibrations, heavy doping, and trap assisted tunneling are analyzed and engineered.

After accounting for the various non-idealities, the ideal density of states will determine the on state characteristics. The nature of the quantum density of states is strongly dependent on dimensionality. Hence we need to specify both the n-side and the p-side dimensionality of pn junctions. For instance, we find that a typical bulk 3d-3d tunneling pn junction has only a quadratic turn-on function, while a pn junction consisting of two overlapping quantum wells (2d-2d) would have the preferred step function response. Quantum confinement on each side of a pn junction has the added benefit of significantly increasing the on-state tunnel conductance at the turn-on threshold. We analytically demonstrate these effects and then give a numerical non-equilibrium greens function (NEGF) model to verify the key results. Finally we introduce some new device designs that will take advantage of the benefits of 2d-2d tunneling.

Table of Contents

Acknowledgments

First and foremost I would like to thank my advisor, Professor Eli Yablonovitch. It truly has been an honor to work with such a brilliant scientist. He has supported me throughout my PhD, always willing to answer my questions and help me understand the most complicated physical challenges by extracting the key physical insights behind any problem. He gave me the freedom to explore the problems that interested me most, while still giving me enough guidance to help me make some great discoveries. It is only with his help and advice that I have come so far and become the engineer that I am.

I would also like thank the other professors that I have had the opportunity to work with. I really appreciate the advice and support that I received from Professor Hu and Professor King during our STEEP and Theme 1 meetings. I would like to thank Professor Salahuddin for helpful theoretical discussions and for posing challenging questions that have helped me refine and improve my dimensionality analyses. I am also very grateful for the opportunity to work with Professor Hoyt and Antoniadis at MIT. Their feedback has been helpful in clarifying my ideas and focusing on the important problems. I would also like to thank Professor Javey and Junqiao Wu for serving on my qualifying exam committee and challenging me to improve my research.

I am also very thankful to Dr. Josephine Yuen for running the E3S center and providing me with many opportunities to develop my professional skills and for advising me on my professional growth. I am also very grateful for Janny Peng's and Dr. Sharnnia Artis's help and support with the E3S Center.

I am also indebted to my amazing colleagues in the device and optoelectronics groups. It is their friendship and support that have allowed me to really grow and that have made the past five years an amazing experience. I am also grateful to Alex Mutig and Jared Carter for closely working with me to experimentally demonstrate the new dimensionality effects that we discovered.

I would also like to thank the NSF and DARPA for funding my education and research. The Center for Energy Efficient Electronics Sciences, which receives support from the National Science Foundation (NSF award number ECCS-0939514), has been a great boon for my research as it has given me the ability to collaborate with a wide range of researchers and learn from the best. I am also grateful for the NSF Graduate fellowship and for DARPA's STEEP program that funded my first few years.

Last but not least, I am grateful to my family for supporting me throughout my education and always pushing me to succeed.

Chapter 1: Introduction

1.1 THE NEED FOR LOW POWER ELECTRONICS

Power consumption is increasingly critical for modern electronics. Reducing the power consumption of electronics can make a significant impact on the worldwide energy demands. In 2010 data centers alone consumed about 2% of all the electricity in the United states as seen in Figure 1.1 [1]. Reducing the power consumption is also critical for portable electronics such as

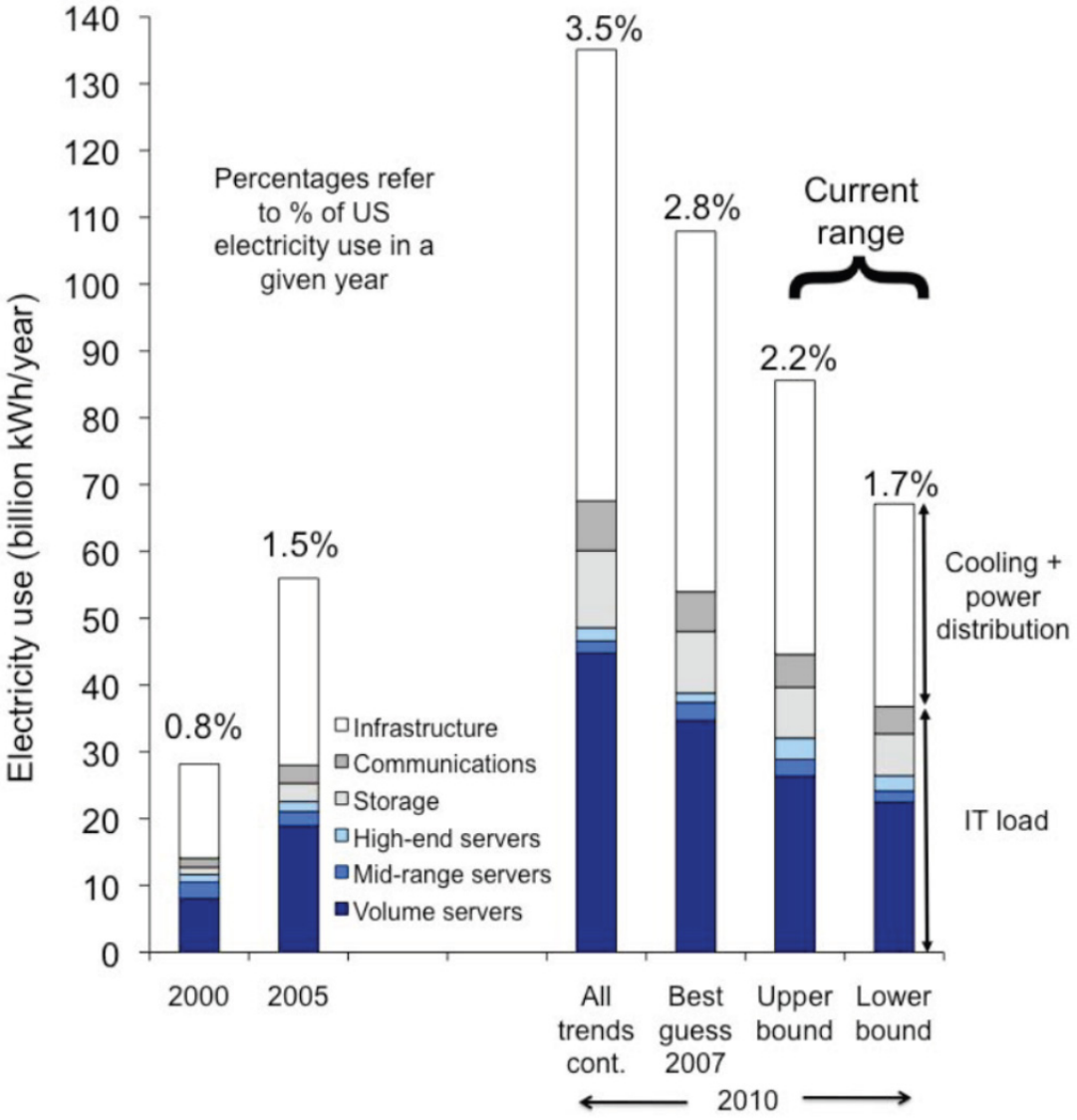


Figure 1.1: US electricity use for data centers (from Koomey 2011)

smartphones whose battery can barely last for a single day. In the past, transistor voltage reduced with shrinking size, but in recent years the voltage scaling has stopped as seen in Table 1.1. At the end of the transistor roadmap [2], the high performance operating voltage is projected to be 0.57 V. Consequently, there has been a growing focus on increasing the number of cores on a chip, even if it means decreasing the clock frequency. This is because the power dissipation increases significantly for a small increase in clock frequency.

Technology Node	0.25 µm	0.18 µm	0.13 µm	90 nm	65 nm	45 nm	32 nm	22 nm	15 nm	10.9 nm
Vdd	2.5 V	1.8 V	1.3 V	1.2 V	1.1 V	1.0 V	0.97 V	0.9 V	0.8 V	0.71 V

Table 1.1: Vdd scaling has slowed after 0.13µm node (From the High Performance ITRS roadmap).

1.2 THE LIMIT OF CURRENT TRANSISTORS

The reason the voltage has stopped scaling is because conventional transistors rely on the thermal excitation of carriers over an energy barrier as shown in Figure 1.2. The probability distribution of electrons follows a Boltzmann factor and at best the current will change by a factor of e for a change in gate voltage of K_bT. This corresponds to a subthreshold slope limit of 60 mV/decade at room temperature. At least 60 mV of bias on the gate is required for each decade of current. While maintaining a good on/off ratio of around 6 orders of magnitude, it is impossible to significantly reduce the supply voltage.

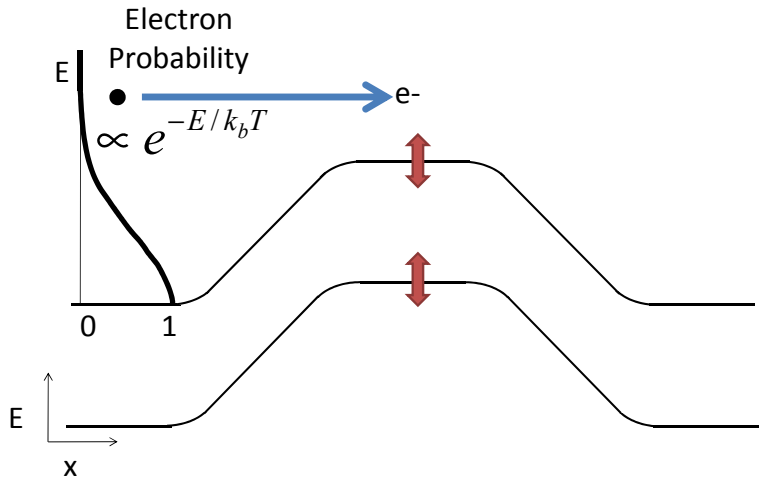


Figure 1.2: Conventional transistors rely on the thermal excitation of carriers over a barrier. If the barrier is shifted by K_bT, the current only changes by a factor of e.

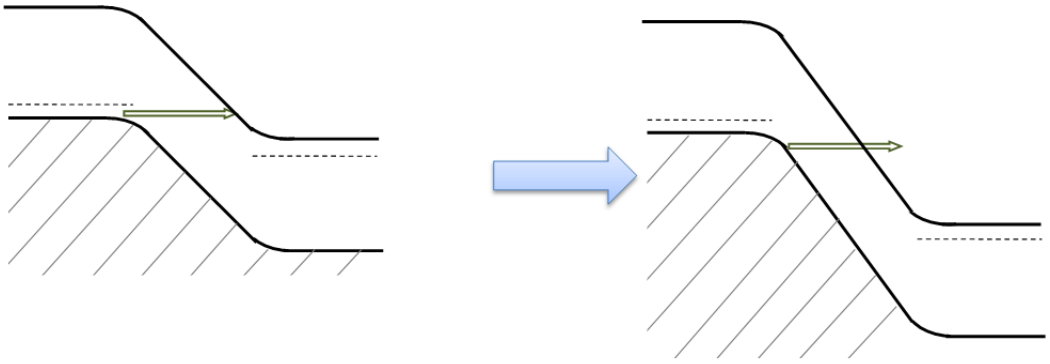


Figure 1.3: The thickness of the tunneling barrier can be controlled to switch a tunneling junction on or off. Applying a bias on the gate changes the electric field in the tunneling barrier and thus the barrier thickness.

1.3 USING TUNNELING FETS FOR LOW POWER

In order to overcome these fundamental limits a new more sensitive switching mechanism is needed. Since we cannot have current going over a barrier, we can have it go through the barrier. The family of Tunneling Field Effect Transistors (TFETs) includes a number of different devices that may be promising for low voltage operation.

When trying to achieve a very sharp TFET turn on there are at least two mechanisms that can be exploited. The gate voltage can be used to modulate the tunneling barrier thickness and thus the tunneling probability [3-6]. This is illustrated in Figure 1.3. The thickness of the tunneling barrier can be controlled by applying a bias to the gate to change the electric field in the tunneling junction. The difficulty in using this method is that at high conductivities there is already a large electric field across the tunneling junction and so the gate bias cannot control the depletion width resulting in a poor subthreshold slope at high conductivities.

Alternatively, it is also possible use the band overlap or density of states turn-on. The

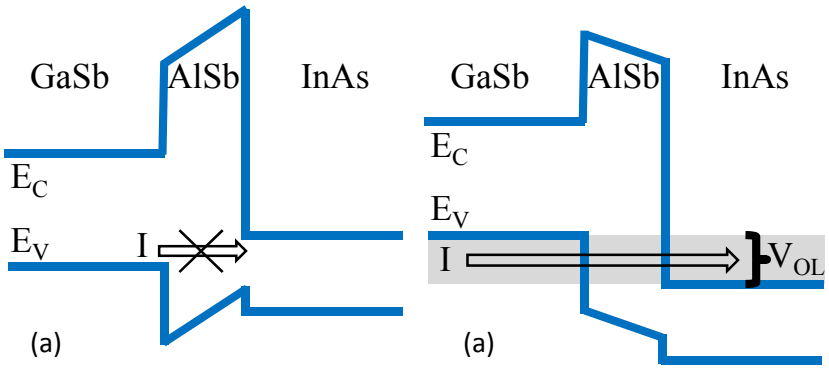


Figure 1.4: (a) No current can flow when the bands do not overlap. (b) Once the bands overlap, current can flow. The band edges need to be very sharp, but density of states arising from dimensionality also plays a role.

band overlap turn-on is illustrated in Figure 1.4. If the conduction and valence band do not overlap, no current can flow. Once they do overlap, there is a path for current to flow. This band overlap turn-on has the potential for a very sharp On/Off transition that is much sharper than that which can be achieved by modulating the tunneling barrier height or thickness[7]. If the band edges are ideal, one might expect an infinitely sharp turn on when the band edges overlap. However, the steepness of the turn on will depend on the density of states of the band edge which will depend on both the ideal density of states as well as various non-idealities such as phonons and dopant or defect induced disorder.

Simulations tend to predict phenomenal TFET performance while experimental results tend to fall short. This can be seen in a review article by Alan Seabaugh [8]. The reason for this is that all the simulations attempt to use the density of states turn on, but do not properly account for the band edge density of states. Nevertheless, there are many experimental results that show sub 60 mV/decade subthreshold slopes at low currents [3-6, 9-11]. However, these devices typically rely on barrier thickness modulation and consequently cannot get a steep subthreshold slope at higher currents. By understanding the physics and limitations of each mechanism we can engineer a new transistor that will overcome these challenges and potentially replace the transistor.

1.4 USING STEEP TUNNEL JUNCTIONS AS A BACKWARD DIODE

Even before creating a full transistor, creating a diode with a sharp turn on near zero bias will be very useful for radio mixing and detection [12, 13]. Backward diodes are tunneling diodes where the tunneling occurs near zero bias as shown in Figure 1.5. The reverse tunneling current results in a highly nonlinear I-V characteristic at small bias. These are used to make high performance detectors and there have already been many interesting devices that perform better than the thermally limited schottky diodes [14-20]. Any improvements to the tunneling junction that results in a sharper turn-on can be immediately used for creating new backward diodes.

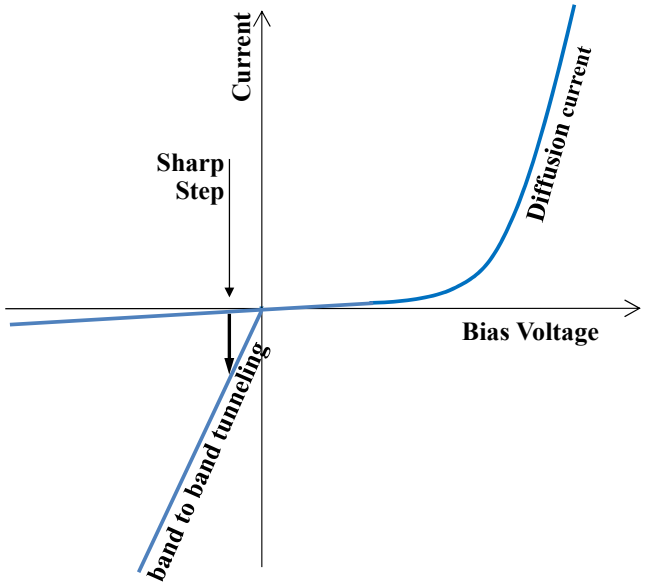


Figure 1.5: Band to band tunneling occurs near zero bias in a backward diode

1.5 RESEARCH OUTLINE

In this dissertation, we will first analyze the limit of tunneling barrier width modulation and show that while it might be interesting for low current applications; it cannot meet all of the desired performance goals. Next, we will analyze the effects of thermal vibrations on the subthreshold slope and propose a few structures to minimize those effects. Then we analyze optical and electrical methods to infer the band edge density of states and show that heavy doping can create an extremely gradual tail of states extending into the band gap. We also model various other limitations such as trap assisted tunneling that can limit a TFET's performance.

After considering and engineering all of the non-idealities in a TFET we will consider the effect of the ideal density of states on the on-state characteristics. The nature of the quantum density of states is strongly dependent on dimensionality. Hence we need to specify both the n-side and the p-side dimensionality of pn junctions. We will find that a typical bulk 3d-3d tunneling pn junction has only a quadratic turn-on function, while a pn junction consisting of two overlapping quantum wells (2d-2d) would have the preferred step function response. Quantum confinement on each side of a pn junction has the added benefit of significantly increasing the on-state tunnel conductance at the turn-on threshold. After analytically demonstrating these effects, we will use a numerical non-equilibrium greens function (NEGF) model to verify the key results. Finally we will introduce some new device designs that will take advantage of the benefits of 2d-2d tunneling.

Chapter 2: Tunnel Barrier Width Modulation

2.1 INTRODUCTION

Controlling the thickness of the tunneling barrier as shown in Figure 1.4 results in steep subthreshold slopes at low current densities and is likely the mechanism behind the best experimental results. This can be seen in the germanium source device in [3]. The experimental I_D-V_G curve closely follows a tunneling barrier width modulation model that is based on gate induced drain leakage. The relevant curves and device structure are reproduced in Figure 2.1. We will examine that model in more detail in the next section. Another steep subthreshold device where the experimental results were matched to a barrier width modulation model is given in [5]. In this paper the simulator Medici was used to model the tunneling current. The tunneling model in Medici does not account for band edges and only considers the thickness of the tunneling barrier. On the contrary, the newer tunneling model in Sentaurus could not fit the experimental characteristics. This is because the model in Sentaurus assumes an abrupt band edge and prematurely cuts off the tunneling current, giving an artificially steep turn-off.

In fact, almost all of the interesting experimental results with a subthreshold slope below 60 mV/decade show tunneling barrier width modulation characteristics. This is indicated by the shape of the I_D-V_G curve that is steep at very low current densities and rapidly rolls off at the higher current densities.

Unfortunately, all of the steep experimental results occur at very low current densities. This is a fundamental limitation of relying on the modulation of the tunneling barrier thickness. As the barrier gets thinner the electric field across it gets larger and it becomes harder to make further changes to the thickness.

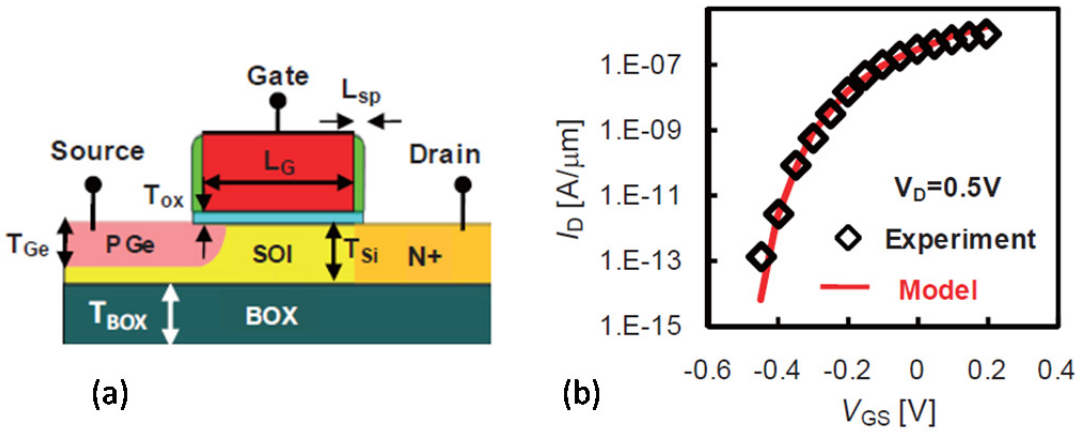


Figure 2.1: (a) schematic of a germanium source TFET (b) The I_D-V_G characteristic closely follows a tunneling barrier width modulation model. (from Kim 2009)

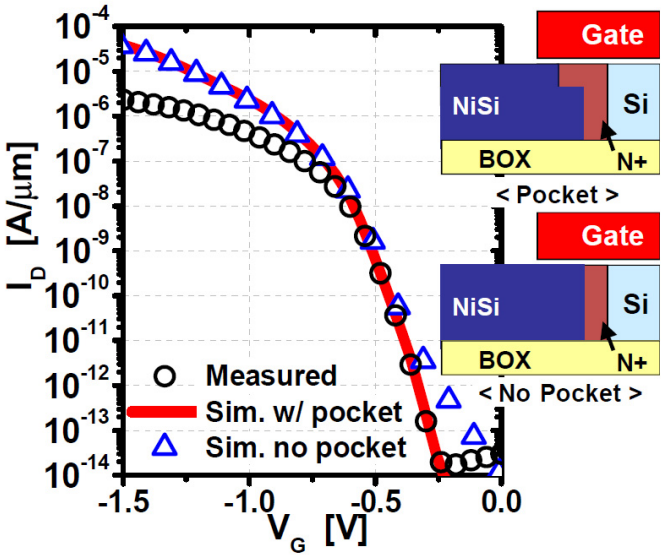


Figure 2.2: The I_D-V_G characteristics of the silicided source TFET shown above fit well with a Medici model that is based on tunneling barrier width modulation. (From Jeon 2011)

2.2 A Simple Barrier Width Modulation Model

In order to examine the effectiveness of modulating the tunneling barrier width we consider the simple model used in [3] that was originally derived to model gate induced drain leakage [21]. The basic concept is that applying a voltage to the gate will change the electric field in the tunneling junction and consequently change its width.

To model the tunneling current we can look at the Kane expression for tunneling[22]. A version of this equation is derived later in Chapter 6:

$$I = \frac{em^* A}{18\hbar^3} \times \frac{\overline{E}_\perp}{2} \times \exp\left(\frac{-\pi (m^*)^{1/2} E_G^{3/2}}{2\sqrt{2}\hbar q\vec{E}}\right) \times \int \left[f_1(E) - f_2(E)\right]\left[1 - \exp(-2E_S / \overline{E}_\perp)\right] dE \qquad (2.2.1)$$

$$\text{where } \overline{E}_\perp = \frac{\sqrt{2}\hbar q\vec{E}}{\pi (m^*)^{1/2} E_G^{1/2}}$$

We see that the dominant effect of the gate bias is to change the tunneling probability which is given by the exponential. We assume the rest of the equation varies slowly and can be taken as a constant. To have a reasonable on current of 100 μA/μm at 1V we will find that we need a tunneling probability of roughly 1%. Given that, we can focus on the exponential and see what the subthreshold slope at a tunneling probability of 1% is. For the moment we will ignore the fact that germanium has an indirect gap. Requiring a phonon to participate can reduce the current by up to three orders of magnitude [22]. Fortunately it may be possible to use dopants or other impurities to relax the requirement for phonons [23]

To estimate the required tunneling probability, we consider the material parameters of germanium. For the prefactor we use a reduced density of states mass of $(1/0.34 + 1/0.22)^{-1} = 0.13$ and a band gap of 0.67eV. We also note that the energy integral is roughly given by the applied bias of 1V and is 1eV. We also assume a tunnel junction size of 1 μm by 10 nm and that the electric field is roughly 0.2 V/ nm. Plugging these into Eqn. (2.2.1) solving for a current of 100 μA gives a required tunneling probability of 1%.

Now that we know the tunneling probability should be around 1% we can focus on the equation for the tunneling probability:

$$T = \exp\left(\frac{-\pi(m^*)^{1/2} E_G^{3/2}}{2\sqrt{2}\hbar q\vec{E}}\right) \tag{2.2.2}$$

The subthreshold slope will be given by the change in the tunneling probability with gate bias:

$$SS = 1\Big/\frac{d\log(T)}{dV_G} \tag{2.2.3}$$

To relate T to V_G we need to calculate how the electric field across the junction changes with V_G. This depends on the exact geometry and TFET design being analyzed. As an example we consider the vertical tunneling junction in Figure 2.1(a). Near the tunneling turn on the channel will be inverted and so the electric field across the oxide will simply be given by:

$$\vec{E}_{OX} = (V_G - V_T)/T_{OX} \tag{2.2.4}$$

The maximum electric field in the semiconductor will occur at the oxide semiconductor interface and is given by the boundary condition $\varepsilon_{OX}\vec{E}_{OX} = \varepsilon_S\vec{E}_S$. This means that the maximum electric field in the semiconductor is given by:

$$\vec{E}_S = \frac{\varepsilon_{OX}}{\varepsilon_S}\times\frac{V_G - V_T}{T_{OX}} \tag{2.2.5}$$

Using this in Eqn. (2.2.2) gives the curve in Figure 2.1(b). Evaluating the subthreshold slope Eqn. (2.2.3) at a given tunneling probability using Eqn (2.2.2) and Eqn (2.2.5) gives

$$SS = \frac{\log(e)}{\log(T)^2}\times\frac{\pi(m^*)^{1/2}E_G^{3/2}}{2\sqrt{2}\hbar q}\times\frac{T_{OX}\varepsilon_S}{\varepsilon_{OX}} \tag{2.2.6}$$

Now we can evaluate this for germanium to see what the subthreshold slope at a tunneling probability of 1% (which corresponds to the desired conductivity of 100 μS/μm at 1V). For the tunneling effective mass we use the geometric mean of the transverse electron mass and the light hole mass as those are the most favorable masses for tunneling and get m^*=0.06 [24]. We also assume an effective oxide thickness of 1 nm. Using these values we get a subthreshold slope of 240 mV/decade at 100 μS/μm. The conductivity at a subthreshold slope of 60 mV/decade is 1 μS/μm. This illustrates the fundamental limitation of barrier width modulation. The steeper subthreshold slopes only occur at the lower current densities. Nevertheless, observing this in practice would be extremely interesting, but unfortunately the indirect gap reduces the observed

current by 2-3 orders of magnitude and the approximations used tend to overestimate the conductivity at a given subthreshold slope.

If we had a direct gap semiconductor, we could ask what material parameters are needed to achieve a slope of 60 mV/decade at a tunneling probability of 1%. Looking at Eqn. (2.2.6) we can see that reducing the effective mass, band gap and permittivity will all reduce the subthreshold slope at a given conductivity. If we only vary the band gap we would need a band gap of 0.26 eV. However, up to this point we have ignored the fact that there is also a mismatch between the conduction and valence band wavefunctions. If we account for this the current would reduce by roughly another order of magnitude meaning that we would need tunneling probability of roughly 10% for a current of 100 μS/μm. At T=10% the barrier height would have to be 105 mV and the barrier would be about 50 nm thick in the off state!

While this is even smaller than the band gap of InAs (0.354 eV) it is possible to achieve an effective tunneling barrier much smaller than this by using a heterojunction. Taking this to its ultimate limit, we could design a barrier modulation heterojunction TFET with a barrier height of only a few mV as is done in the next section. Unfortunately, we will see that the maximum on-off ratio will be limited by the fact that the band edges are not perfectly sharp, but rather have a density of states tail that extends into the band gap and so the electrons will never see the extremely thick barriers that are required in the off state.

2.3 THE ULTIMATE LIMITS OF BARRIER WIDTH MODULATION

By taking barrier width modulation to its ultimate limit, we could design a barrier modulation heterojunction TFET with a barrier height of only a few mV. If we assumed perfectly sharp band edges we find that the device can be optimized to provide a very steep swing, on the order of a few millivolts per decade, at a high current density over about a decade of current or it can be optimized to provide a slightly worse swing over several decades of current at a lower current density. Unfortunately, as we will see in the following section, accounting for the band tails will prevent the following proposal from operating as intended.

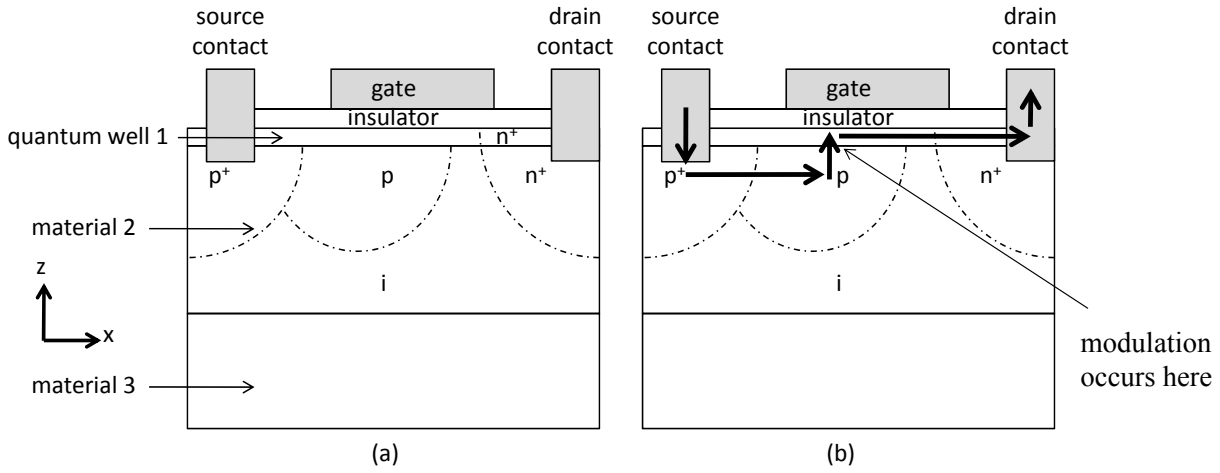


Figure 2.3: (a) Possible implementation of new gate control mechanism (b) structure with the current path labeled

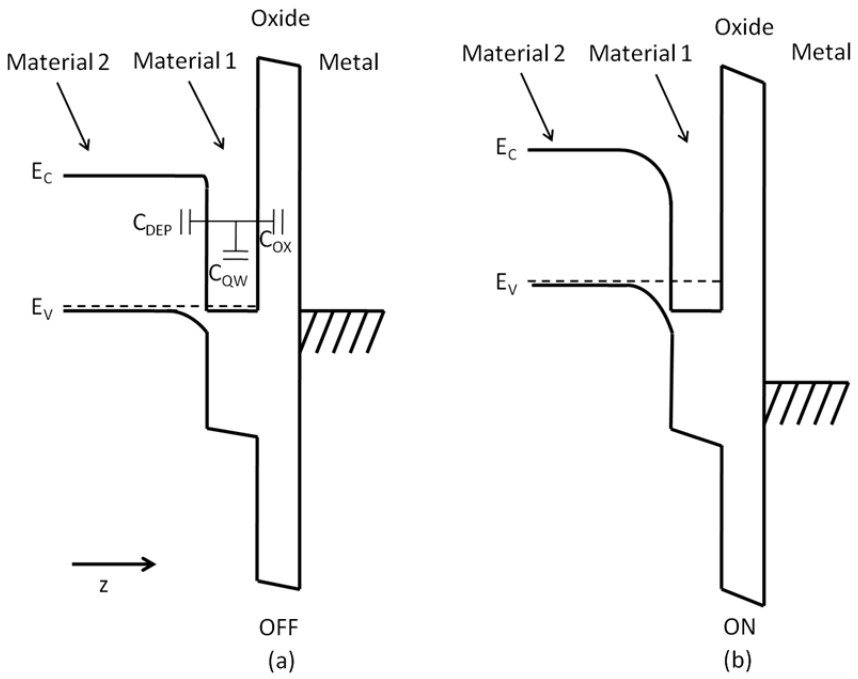


Figure 2.4: Band diagram of the proposed device in the (a) ON and (b) OFF states

The proposed device is shown in Figure 2.3a, with the current path shown in Figure 2.3b. The vertical band diagram under the gate (along the z-axis) shown in Figure 2.4. Figure 2.4a shows the band diagram of the device right before it turns on. This is a vertical tunneling device with a quantum well on top of a strained layer. A blow up of the region where the tunneling occurs is shown in Figure 2.5. The tunneling barrier, Δ, is set by the heterostructure band alignment between materials 1 and 2. By choosing the correct materials, the barrier can be made very small to allow a large amount of current to pass or it can be made larger to have a steeper response over more decades of current, but at a lower current density. One possible combination of materials is InAs for material 1 and GaSb or GaAlSb for material 2. Introducing Al into the GaSb allows the barrier height to be fine-tuned. While InAs/GaSb form a type 3 heterostructure, the InAs (Material 1) will be a quantum well and so the confinement energy will effectively create the type 2 heterostructure shown in Figure 2.4 and Figure 2.5. The quantum well size can also be used to fine tune the barrier height through the confinement energy. Any material system that creates a type 2 or type 3 heterostructure can be used, and since material 1 will be a thin quantum well the materials do not have to be lattice matched. Material 1 can even be a conducting layer of interface states. In Figure 2.3, materials 2 and 3 can be the same. However, material 3 can be chosen to strain material 2 such that the heavy hole band is raised above the light hole band. This improves the subthreshold slope, as will be shown later. One possible option for material 3 is AlSb or GaAlSb if material 2 is GaSb.

As shown in Figure 2.4, when an increasing gate bias is applied, more electrons are added to the 2d channel and so the splitting between the Fermi level and the conduction band increases. Ideally, most of the change in gate voltage should be transmitted to material 2. In order to minimize the potential change across the quantum well and maintain the same the relative carrier distribution in the channel, the quantum well needs to be very narrow and on the order of a nanometer. Furthermore simply using a quantum well in the channel reduces the capacitance

and thus improves the gate coupling. Since the channel quantum well is very thin the doping in the channel will not have significant effect and the channel potential will be set by the gate. Consequently the channel doping can be arbitrarily set. At most the channel doping will shift the threshold voltage.

As seen in Figure 2.5, the position of the Fermi level is fixed with respect to the bulk valence band in material 2. With an increasing gate voltage the potential drop across material 2 increases and since the band offsets are rigidly fixed, the valence band must curve downwards and so the tunneling barrier height with respect to the Fermi level must increase. However since the offset between the conduction band in material 1 and the valence band in material 2 is fixed at Δ, the tunneling barrier height for states at the bottom of the conduction band is fixed. Nevertheless, the barrier width will decrease as the electric field near the junction increases with increasing gate voltage. This is shown in Figure 2.5 where w' is less than w. This means that the current will increase as the gate voltage increases. While it is true that due to Fermi-Dirac statistics, the available states to tunnel to decreases as the conduction band gets farther away from the Fermi level, it takes a 25 mV change (k_BT) in the energy level to reduce the available states by a factor of e. However, a few mV change in the conduction band can significantly decrease the tunneling barrier and thus significantly increase the tunneling probability.

The heavily doped p+ and n+ regions in Figure 2.3 are used to make the source and drain contacts. The source contact simply needs to contact the p-region in material 2 and any method can be used to make that contact. Likewise the drain contact simply needs to contact material 1. However, the n+ doping of the drain contact should not come in contact with the p doped region or else it is possible for a leakage path to form. Similarly the n+ doping should extend to the gate to reduce the series resistance.

2.3.1 Quantitative Analysis

2.3.1.1 Gate Coupling Efficiency

First we consider the coupling between the gate and the depletion region. Ideally, most of the voltage should be transmitted through to the depletion region and so a small quantum well thickness and small equivalent oxide thickness (EOT) are desired. In order to estimate the

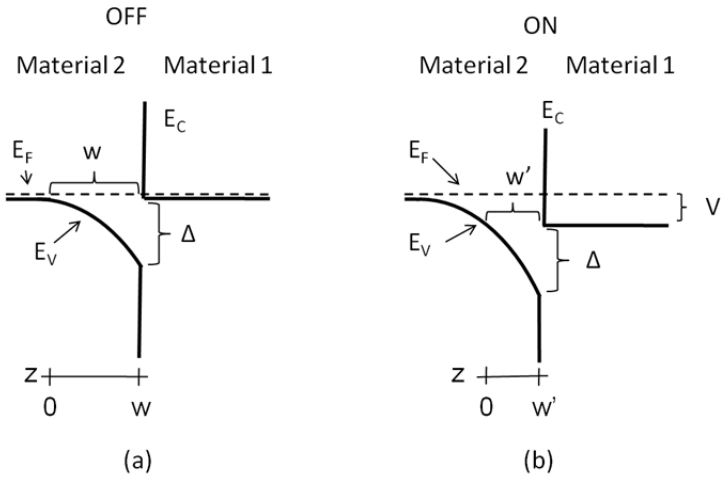


Figure 2.5: Tunneling portion of the band diagram

voltage on material 2 we will make approximations that give a conservative worst case estimate of the gate coupling. The oxide, quantum well channel and depletion region in material 2 can be modeled as three capacitors which are labeled by C_{OX}, C_{QW}, and C_{DEP}, respectively. The channel needs to be a quantum well in order to minimize the capacitance. The quantum well channel is in parallel with the depletion region and both of them are in series with the gate oxide as shown in Figure 2.4a. Thus the change in the surface potential of the semiconductor at the oxide interface, V_s, is given by:

$$V_s = \frac{C_{OX}}{C_{OX} + C_{QW} + C_{DEP}} V_g \tag{2.3.1}$$

Where

$$C_{OX} = \varepsilon_{OX} / t_{OX} \tag{2.3.2}$$

$$C_{QW} = q^2 \frac{dN}{dE} = \frac{q^2 m^*}{\pi \hbar^2} \tag{2.3.3}$$

$$C_{DEP} = \sqrt{\frac{qN_a \varepsilon_2}{2V_{DEP}}} \tag{2.3.4}$$

V_g is the voltage applied on the gate relative to the threshold voltage. In order to estimate this, we consider material 1 to be InAs and material 2 to be GaSb and consider an EOT of 1 nm with $\varepsilon_{OX}=3.9\varepsilon_0$. The electron effective mass for InAs, m^*, is taken to be 0.023, the relative permittivity for GaSb, ε_2, is $15\varepsilon_0$ and a doping level, N_a, of $10^{16}/cm^3$ is chosen. The voltage across the depletion region, V_{DEP}, is chosen to be 10 mV. This gives the following capacitances: $C_{OX}=3.45*10^{-6}$ F/cm^2, $C_{QW}=1.54*10^{-6}$ F/cm^2, $C_{DEP}=3.26*10^{-7}$ F/cm^2. Thus $V_s= 0.65\ V_g$.

The quantum well and the depletion region are not exactly in parallel as some voltage is lost across the quantum well before the gate potential reaches the depletion region in material 2. A worst case estimate of this voltage is the peak field in the quantum well times the well width. The peak field is the field set by the oxide. The voltage dropped across the quantum well will be:

$$\Delta V_{QW} = \vec{E}_{QW} * t_{QW} = \frac{\varepsilon_{OX}}{\varepsilon_{QW}} \frac{V_{OX}}{t_{OX}} * t_{QW} = \frac{\varepsilon_{OX}}{\varepsilon_{QW}} \frac{t_{QW}}{t_{OX}} * 0.35 * V_G \tag{2.3.5}$$

V_{OX} is the voltage across the oxide. If a quantum well thickness, t_{QW}, of 1nm is assumed and $\varepsilon_{QW}= \varepsilon_{InAs}=15\varepsilon_0$, the voltage lost across the quantum well is less than 9% of the gate voltage. Thus the voltage change in the depletion region is $V=(0.65-0.09)*V_g=0.56*V_g$. Over half of the gate voltage is transmitted to the depletion region.

2.3.1.2 Tunneling Probability and Subthreshold Slope

There are two mechanisms that can turn the device on and off. First the conduction and valence band need to overlap in order for there to be states available for tunneling. At first it seems like there will be a sudden transition as the gate bias is increased where current is allowed to flow and so this process can result in a very steep subthreshold slope as. However, the band edges are not necessarily very sharp and there will be band tails that will limit the subthreshold slope. Nevertheless, this process will still be better than the current thermally limited

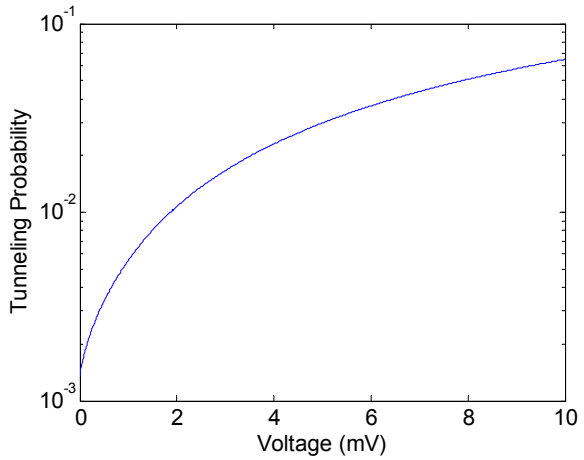


Figure 2.6: Tunneling probability for a 5 mV barrier

subthreshold slope of 60mV/decade. As such, this device can be optimized for this process by making the tunneling barrier width as thin as possible by doping material 2 heavily.

The tunneling current can also be modulated by adjusting the depletion region width through the gate bias. In the following analysis we consider this effect in order to obtain a subthreshold slope that is even steeper than that which can be obtained from the band overlap effect alone. Consequently we ignore the band overlap effect and consider only the tunneling probability.

The tunneling probability can be estimated from the WKB approximation. The tunneling barrier is shown in Figure 2.5. Electrons tunnel across a parabolic barrier from the valence band to the conduction band. The states with the greatest tunneling probability (T) are those at the band edge and so we will consider those states. Let V' be the barrier height at a given position z. Then we have:

$$T \propto e^{-2\int_0^{w'} k \cdot dz} \tag{2.3.6}$$

Where $k = \sqrt{2m^*(V'-V)/\hbar^2}$ and $z = \sqrt{2\varepsilon V'/qN_a}$. Thus

$$T \propto e^{-\frac{1}{V_s}*\int_V^{V+\Delta} \sqrt{\frac{V'-V}{V'}}\partial V'} \tag{2.3.7}$$

$$T \propto \left(\frac{\sqrt{V+\Delta}+\sqrt{\Delta}}{\sqrt{V}}\right)^{V/Vs} * e^{-\frac{\sqrt{\Delta(V+\Delta)}}{V_s}} \tag{2.3.8}$$

Where

$$V_s = \frac{\hbar}{2}\sqrt{\frac{N_a}{m^*\varepsilon}} \tag{2.3.9}$$

The voltage V is the voltage that is transmitted to the depletion region and is about half the applied gate voltage. Δ is the built in barrier set by the heterostructure band alignment. m^* and ε

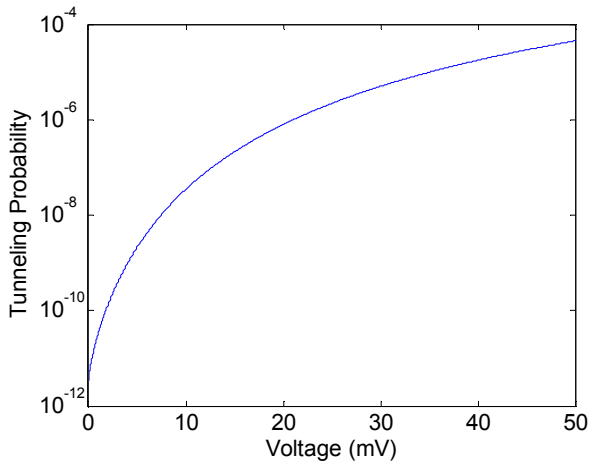


Figure 2.7: Tunneling probability for a 20 mV barrier

are the effective mass and permittivity, respectively, in material 2. V_s is a parameter that sets the steepness of the subthreshold slope. The smaller V_s is the steeper the subthreshold slope will be. This implies that a light doping and heavy effective mass are desired. To estimate the optimal device performance we consider low doping level of $10^{16}/cm^3$ and the heavy hole effective mass of GaSb of $0.4m_0$. This gives a Vs of 0.76mV. Given this value of Vs, the tunneling probability as a function of V is plotted in Figure 2.6 for a barrier height, Δ, of 5 mV and for a barrier height of 20 mV in Figure 2.7. As seen in Figure 2.6, a 2 mV change changes the tunneling probability by a decade and the on state tunneling probability is roughly 1%. After including the gate coupling, this corresponds to a subthreshold slope of roughly 4 mV/decade. Figure 2.7 shows that a change in six decades of current for a 25 mV change in potential is possible with a 20 mV barrier. However, in this case the on state current is significantly reduced and the on state tunneling probability is 10^{-6}.

2.3.2 Conclusion

This device exploits a unique heterostructure and surface quantum well to create a new transistor with an extremely steep subthreshold slope. By varying various parameters such as the doping or the material compositions this device can be optimized for a range of performance metrics. It can provide a steep subthreshold slope over many decades of current at a lower overall current density, or it can provide a very steep subthreshold slope at a high current density, but only for a limited range of current. These different modes of operation mean that the device will have many potential applications.

2.4 So What Went Wrong?

Based on the description in Section 2.3 is seems like TFETs should be solved. However we have not accounted for states that extend into the band gap. As shown in Figure 2.8, the band edge is not perfectly sharp, but rather there is a tail of states extending deep into the band gap. Since we were considering a barrier height of only 20 mV, the electrons will never see the barrier, but rather they will pass directly into the band tail and the device will not turn off. As will be shown in Chapter 4, the intrinsic phonon induced band tails in silicon are around 27

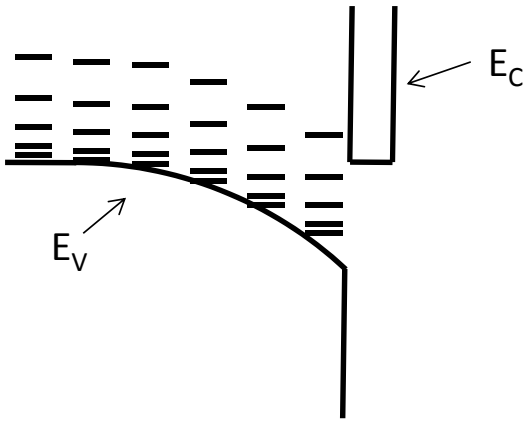


Figure 2.8: Current can easily pass through the band tails, preventing the barrier modulation from working as intended.

mV/decade. This means that with a 20 mV barrier we will see less than 1 decade on to off ratio before the band tails dominate the I-V characteristic.

2.5 Is There Any Hope for Barrier Modulation?

As we saw in Section 2.4, using a very small barrier on the order of 10's of millivolts will not work. However, as we saw at the end of Section 2.2, we should be able to see useful results with a 260 mV high tunneling barrier at lower conductances. Since there will be band tails on both the conduction band and the valence band, the band edge density of states midway through the effective gap will control the off state current. This means we should ask, "What is the density of states 130 mV from the band edge?"

If we assume the junction has been designed well and the only contribution to the band tail is the intrinsic phonon limited tail described in Chapter 4, the density of states will fall off at a rate of 27 mV/decade. Consequently we could have up to 130/27= 4.8 decades of on to off ratio before being limited by the band tails and even then the band tails would provide a slope of 27 mV/decade resulting in a high performance device.

However, if the junction is poorly designed and uses heavy doping, the band tails could be worse than 100 mV/decade as described in Chapter 4. In this case the on to off ratio for barrier modulation would only be 130/100= 1.3 decades of current and then the subsequent band tail turn off would be 100 mV/decade. This would be a very poor device.

Overall, using barrier width modulation could be interesting with a small barrier height, but only if the band tails are properly accounted for. Unlike current barrier width modulation devices, a properly designed one would use barrier width modulation to provide a steeper slope at the higher current densities while the slope at lower current densities would be controlled by the band tail density of states. However, given that the band tail density of states needs to be optimized it might still be better to design a switch that operates exclusively on the density of states overlap. The tradeoff between the different possibilities needs to be analyzed further. As it stands now, it seems that barrier thickness modulation is unlikely to provide the desired performance at the higher current densities.

Chapter 3: Engineering the Deformation Potential Limits

3.1 INTRODUCTION

In practical devices the band edges will not have an ideal density of states that fall sharply to zero at the band edge, but rather there will be a band tail. This tail will be caused by any imperfections in the lattice, whether they are due to impurities or phonons. In optical measurements this results in the Urbach tail of the absorption spectrum. In silicon the optical absorption coefficient falls off as an exponential at the rate of 27 mV/decade [25, 26]. A similar tail exists in tunneling devices and it will pose a similar limit on the achievable sub-threshold slope. This can be seen in some non-equilibrium greens function (NEGF) simulations that account for phonon scattering [27-32]. Once phonons are included the best achievable sub-threshold slope is significantly reduced due to the phonon band tails. Nevertheless, by engineering the electron-phonon interactions it may be possible to reduce the phonon effects. In this chapter we will focus on long wavelength acoustic phonons, how they contribute to the 27 mV/decade limit and how to reduce their effects.

3.2 THERMAL GENERATION OF STRAIN WAVES

Thermal vibrations can be represented as phonons or displacement waves. These phonons will cause random strains that cause the band edge energies to shift. Every point in the first brillouin zone of reciprocal space corresponds to three acoustic phonon modes and three optical phonon modes for typical semiconductors. If we consider a device operating at 10 GHz, the optical phonons oscillate roughly a thousand times faster on the order of 10^{13} Hz[33]. Thus any energy shifts that they cause will be subject to motional narrowing, or time averaging. We have not considered the possibility of directly absorbing an optical phonon.

The three acoustic modes at each point in k-space can be divided into a longitudinal mode and two transverse modes. In this analysis we approximate the phonon dispersion relationship as linear, i.e. $\omega = v_s |\vec{k}|$, where ω is the phonon frequency, k is the phonon wave vector and v_s is the speed of sound for the phonon mode. An arbitrary phonon mode can be written as:

$$\delta\vec{R} = (A_x\hat{x} + A_y\hat{y} + A_z\hat{z})\cos(k_x x + k_y y + k_z z - \omega t) \tag{3.2.1}$$

The strains are also defined as[34]:

$$\begin{aligned} \varepsilon_{xx} &= \frac{\partial \delta R_x}{\partial x} & \varepsilon_{xy} &= \frac{1}{2}\left(\frac{\partial \delta R_x}{\partial y} + \frac{\partial \delta R_y}{\partial x}\right) \\ \varepsilon_{xx} &= \frac{\partial \delta R_y}{\partial y} & \varepsilon_{xz} &= \frac{1}{2}\left(\frac{\partial \delta R_x}{\partial z} + \frac{\partial \delta R_z}{\partial x}\right) \\ \varepsilon_{zz} &= \frac{\partial \delta R_z}{\partial z} & \varepsilon_{yz} &= \frac{1}{2}\left(\frac{\partial \delta R_y}{\partial z} + \frac{\partial \delta R_z}{\partial y}\right) \end{aligned} \tag{3.2.2}$$

Applying this definition (3.2.2) to the phonon wave equation (3.2.1) results in following strain tensor:

$$\begin{aligned} \vec{\varepsilon} = &- |A||k| \sin(\vec{k}\cdot\vec{r} - \omega t)\frac{1}{|A||k|} \\ &\cdot \begin{pmatrix} A_x k_x & A_x k_y + A_y k_x & A_x k_z + A_z k_x \\ A_x k_y + A_y k_x & A_Y k_Y & A_y k_z + A_z k_y \\ A_x k_z + A_z k_x & A_y k_z + A_z k_y & A_Z k_Z \end{pmatrix} \end{aligned} \tag{3.2.3}$$

For longitudinal waves $\vec{k}$ is approximately parallel to $\vec{\mathrm{A}}$ and for transverse waves $\vec{k}$ is approximately perpendicular to $\vec{\mathrm{A}}$.

We will need to know the variance, or the root mean square (RMS) magnitude, of each type of strain in order to engineer the subthreshold slope as shown in the next section. The magnitude of the strain wave, $\varepsilon_o = |A||k|$ can be found be setting the strain energy[35] of each phonon mode equal to the thermal energy of each mode.

$$\begin{aligned} \mathrm{U} = \int_{crystal} &\frac{1}{2} C_{11}(\varepsilon_{xx}^2 + \varepsilon_{yy}^2 + \varepsilon_{zz}^2) + 2C_{44}(\varepsilon_{xy}^2 + \varepsilon_{yz}^2 + \varepsilon_{xz}^2) \\ &+ C_{12}(\varepsilon_{yy}\varepsilon_{zz} + \varepsilon_{zz}\varepsilon_{xx} + \varepsilon_{xx}\varepsilon_{yy})\partial V = \frac{1}{2}k_b T \end{aligned} \tag{3.2.4}$$

where C_{11}, C_{12} and C_{44} are the elastic constants of the material. When working in reciprocal space each phonon mode can create a full tensor of coherent strains. In Section 3.5 (Appendix A) the equations of motion in a solid are solved in order to find the displacement eigenmodes (3.2.1). Those modes are then used to compute the strains (3.2.3) due to each phonon and then the strains are summed over the first brillouin zone. This finally gives the RMS strains ε_{ij}. Fortunately the strains can be found by a simpler and more intuitive method by making a simplifying approximation and working in a 'strain space' where each strain mode, ε_{ij}, is independent. The results in Section 3.5 (Appendix A), which may not be more accurate due to the linear dispersion approximation, are within 5% of the results (3.2.11, 3.2.12) found by the much simpler method that follows.

In a bulk material each of the uniaxial strains $(\varepsilon_{xx}, \varepsilon_{yy}, \varepsilon_{zz})$ will have the same magnitude as each other and each of the shear strains $(\varepsilon_{xy}, \varepsilon_{yz}, \varepsilon_{xz})$ will have the same magnitude by symmetry. An arbitrary strain mode, ε_a, will have the following form:

$$\varepsilon_a = \varepsilon_o \sum_{\vec{k}} \alpha_{a,\vec{k}} \cos(\vec{k}\cdot\vec{r} - \omega_k t + \phi_{a,\vec{k}}) \tag{3.2.5}$$

where $\sum_k \alpha_{a,k}{}^2 = 1$ and $\alpha_{a,k}$ and $\phi_{a,k}$ are chosen to make each strain mode independent and represent a single degree of freedom. Thus for a uniaxial strain we have:

$$\int_{crystal} \frac{1}{2} C_{11} \varepsilon_a{}^2 dv = \frac{1}{4} C_{11} \varepsilon_o{}^2 V_{crystal} = \frac{1}{2} k_b T \tag{3.2.6}$$

$$\varepsilon_o{}^2 = \frac{2k_b T}{C_{11} V_{crystal}} \tag{3.2.7}$$

The total mean squared strain of a uniaxial mode, $<\varepsilon_{ii}{}^2>$, is:

$$<\varepsilon_{ii}{}^2> = \sum_{\text{degrees of freedom}} \frac{2k_b T}{C_{11} V_{crystal}} = \frac{2k_b T}{C_{11} V_{crystal}} * N_{\text{deg}} \tag{3.2.8}$$

where N_{deg} is the number of degrees of freedom. Similarly the total mean squared strain for a shear mode is:

$$<\varepsilon_{ij}{}^2> = \frac{k_b T}{2C_{44} V_{crystal}} * N_{\text{deg}} \tag{3.2.9}$$

The number of degrees of freedom can be found by comparison to the number of phonon modes in reciprocal space. The number of points in reciprocal space is equal to the number of unit cells in reciprocal space, $N_{cells} = V_{crystal} / V_{\text{unit cell}}$. Each point in reciprocal space has three degrees of freedom corresponding to the 3 acoustic phonon modes (the optical phonons have already been neglected). One degree corresponds to the three uniaxial strains, one degree to the three shear strains, and one degree to the three rotations that take the form $\frac{1}{2}\left(\partial \delta R_i / \partial j - \partial \delta R_j / \partial i\right), i \neq j$ and do not affect the energy. Thus each strain mode has:

$$N_{\text{deg}} = \frac{1}{3} N_{cells} = \frac{V_{crystal}}{3V_{\text{unit cell}}} \tag{3.2.10}$$

However, the strains with a short wavelength may average out or those with a high frequency may be subject to motional narrowing. Consequently we will let β equal the fraction of strain modes that contribute to the energy. In general $0 < \beta \leq 1$ and we assume that it will be the same for shear and uniaxial strains. Thus we finally get the following expressions for the mean squared strains:

$$<\varepsilon_{ii}{}^2> = \frac{2k_b T}{3C_{11} V_{\text{unit cell}}} * \beta \tag{3.2.11}$$

$$<\varepsilon_{ij}{}^2> = \frac{k_b T}{6C_{44} V_{\text{unit cell}}} * \beta, \quad i \neq j \tag{3.2.12}$$

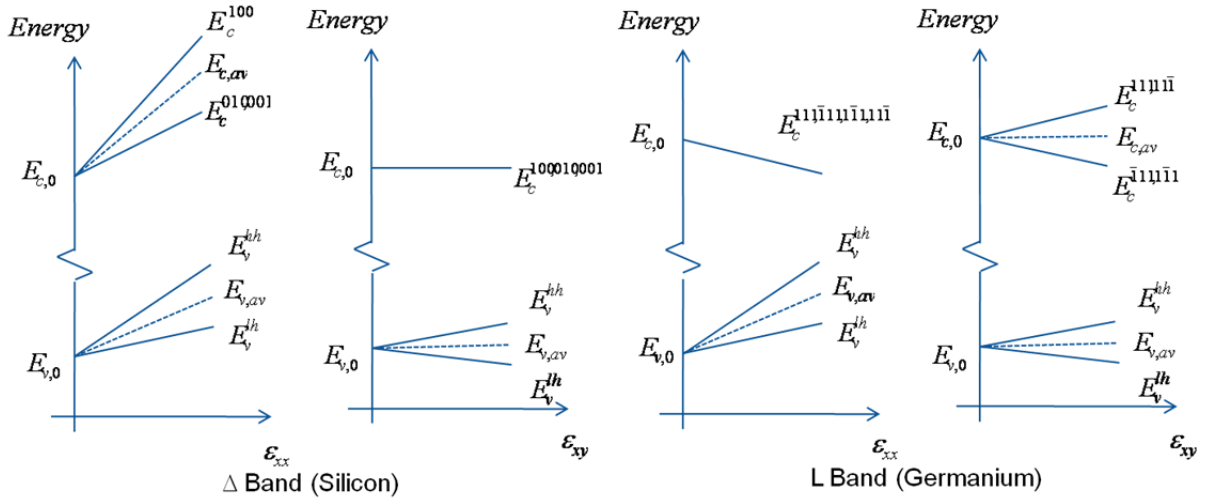


Figure 3.1: Variation of band edges and band splitting as function of strain

3.3 ENERGY SHIFTS DUE TO STRAIN

3.3.1 Conduction and Valance Band Response to Strains

The strain waves cause shifts in the band edge energies. This is qualitatively illustrated in Figure 3.1 (based on [36]). The average conduction band energy shifts are given by[37]:

$$\Delta E_{c,av} = \left(\Xi_d + \frac{1}{3}\Xi_u\right)\left(\varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz}\right) \tag{3.3.1}$$

The constants (Ξ_d, Ξ_u, a, b, and d) are given in Section 3.6 (Appendix B). In addition to the average shift, the degenerate minima also split. The conduction band minima along the <100> directions split in pairs according to the following equations[37]:

$$\begin{aligned}
\Delta E_c^{100} - \Delta E_{c,av} &= \Xi_u^{\Delta}\left(\frac{2}{3}\varepsilon_{xx} - \frac{1}{3}\varepsilon_{yy} - \frac{1}{3}\varepsilon_{zz}\right) \\
\Delta E_c^{010} - \Delta E_{c,av} &= \Xi_u^{\Delta}\left(-\frac{1}{3}\varepsilon_{xx} + \frac{2}{3}\varepsilon_{yy} - \frac{1}{3}\varepsilon_{zz}\right) \\
\Delta E_c^{001} - \Delta E_{c,av} &= \Xi_u^{\Delta}\left(-\frac{1}{3}\varepsilon_{xx} - \frac{1}{3}\varepsilon_{yy} + \frac{2}{3}\varepsilon_{zz}\right)
\end{aligned} \tag{3.3.2}$$

The conduction band minima along the <111> directions split according to the following equations [37]:

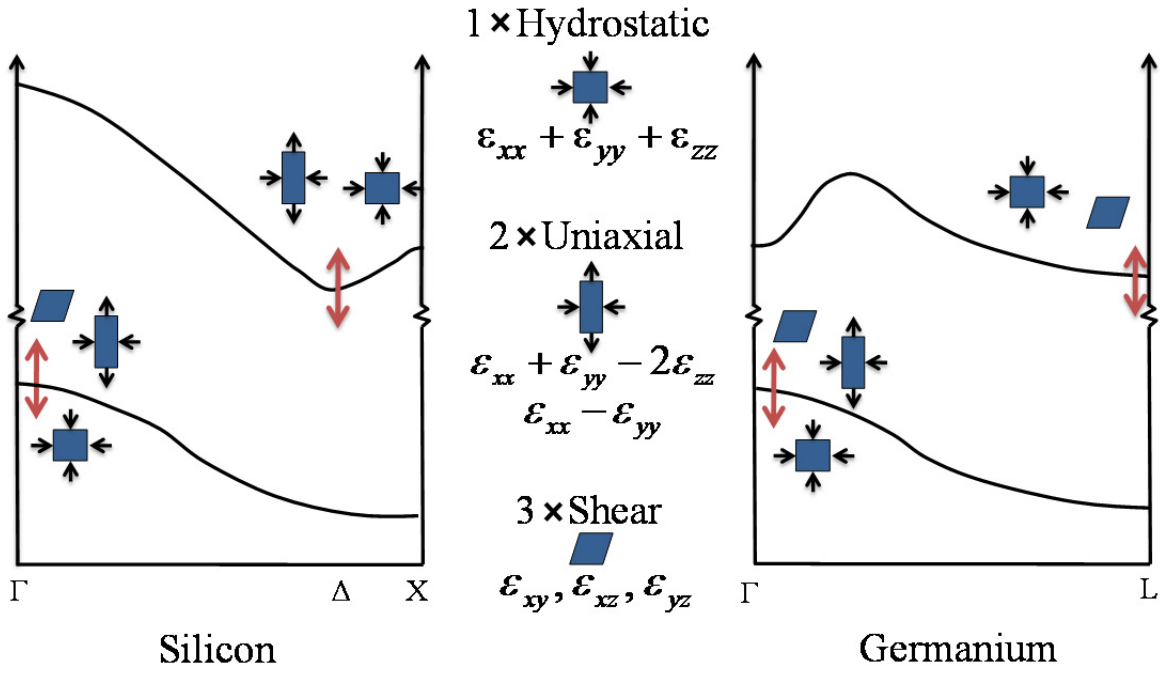


Figure 3.2: Different types of strain cause band extrema to shift

$$\begin{aligned}
\Delta E_c^{111}-\Delta E_{c,av} &= \frac{1}{3}\Xi_u^L\left(\varepsilon_{xy}+\varepsilon_{yz}+\varepsilon_{xz}\right)\\
\Delta E_c^{\bar{1}11}-\Delta E_{c,av} &= \frac{1}{3}\Xi_u^L\left(-\varepsilon_{xy}+\varepsilon_{yz}-\varepsilon_{xz}\right)\\
\Delta E_c^{1\bar{1}1}-\Delta E_{c,av} &= \frac{1}{3}\Xi_u^L\left(-\varepsilon_{xy}-\varepsilon_{yz}+\varepsilon_{xz}\right)\\
\Delta E_c^{11\bar{1}}-\Delta E_{c,av} &= \frac{1}{3}\Xi_u^L\left(\varepsilon_{xy}-\varepsilon_{yz}-\varepsilon_{xz}\right)
\end{aligned} \tag{3.3.3}$$

The valence band energy also shifts with strain and the heavy and light hole band degeneracy is lifted. The energy shifts including quantum confinement along the z [001] direction are shown below [38]. In the unconfined case let, the confined wavevector, k_z=0.

$$\begin{aligned}
E\left(\vec{k},\vec{\varepsilon}\right) = \quad & a\sqrt{3}\varepsilon_1-\frac{k_z{}^2\hbar^2\gamma_1}{2m_0}\\
& \pm\sqrt{\left(\frac{k_z{}^2\hbar^2\gamma_2}{m_0}+\frac{\sqrt{6}}{2}b\varepsilon_2\right)^2+\left(\frac{\sqrt{6}}{2}b\varepsilon_3\right)^2+d^2\left[\varepsilon_{xy}^2+\varepsilon_{yz}^2+\varepsilon_{xz}^2\right]}
\end{aligned} \tag{3.3.4}$$

where:

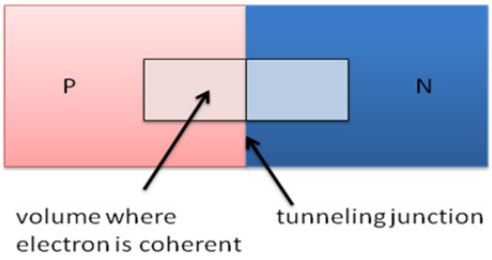


Figure 3.3: Tunneling junction showing region where electron is coherent

$$\varepsilon_1 = \frac{1}{\sqrt{3}}\left(\varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz}\right)$$
$$\varepsilon_2 = \frac{1}{\sqrt{6}}\left(\varepsilon_{xx} + \varepsilon_{yy} - 2\varepsilon_{zz}\right) \qquad (3.3.5)$$
$$\varepsilon_3 = \frac{1}{\sqrt{2}}\left(\varepsilon_{xx} - \varepsilon_{yy}\right)$$

where γ_1 and γ_2 are the luttinger parameters and m_0 is the free electron mass. The different types of strain that affect each of the band extrema are shown in Figure 3.2.

3.3.2 Engineering the Energy Shifts

The tunneling junction between a p-doped semiconductor and a n-doped semiconductor is shown in Figure 3.3. Throughout the semiconductor, the band edge energy will vary randomly due to the strains. However an electron tunneling from one side to another will approximately respond to the average energy over the volume in which the electron is coherent. This means that if the strain wavelength is considerably shorter than the coherence length, the effects of the short wavelength strains will average out. To get an order of magnitude estimate of the coherence volume, consider the wavelength of an electron with k_bT of energy. At room temperature for silicon this is roughly 10 nm ($k_bT = \hbar^2k^2/2m^*$). This means that strains with wavelengths considerably shorter than 10 nm will average out and have no effect, reducing the factor β in Eqs. (12)-(13). High frequency strains or phonons will also tend to average out due to motional narrowing. Strains faster than the device, on the order of 100Ghz and higher, which correspond to wavelengths of 50 nm and shorter[33] in silicon, will likely be subject to motional narrowing. This means that primarily strains with wavelengths greater than 50 nm will contribute to the energy level shifts. Since the strain wavelengths are much larger than the coherence length, the energy shift due to any one phonon will be roughly constant over the entire coherence volume. Consequently any given electron will see an approximately uniform thermal strain over its entire coherence volume, which includes both the P and N sides of the tunneling junction. The direct quantum mechanical absorption and emission of phonons was not considered. If this allows shorter wavelength phonons to contribute, the following methods can be considered a partial solution to phonon problem.

3.3.2.1 Response of Bulk Valence Band to Thermal Strains

The degeneracy between the heavy hole and light hole bands is broken whenever a strain is present. However, in a bulk band, thermal strains are still present and so it seems like the bands should split. However, this is not observed in practice. Motional narrowing will reduce

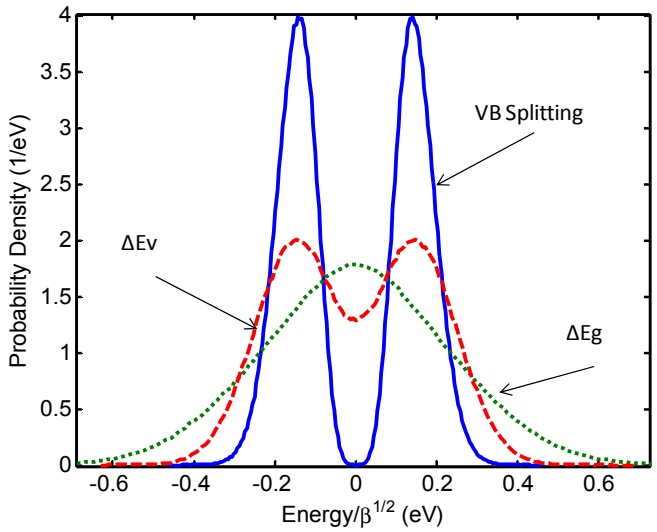


Figure 3.4: Probability distributions of the energy shifts. Solid line: distribution of the valence band splitting. Dashed line: distribution of the valence band energy. Dotted line: distribution of the band gap energy.

the effect, but it alone is not sufficient to explain why the splitting is not observed. In order to get a better understanding of what is happening we will assume that each of the strains has a Gaussian distribution centered at zero. This is reasonable as each strain component is the sum of many random degrees of freedom. The distribution of the energy splitting or square root term in (3.3.4) is related to a chi distribution and is the solid line in Figure 3.4. The plot is normalized to have an area of 1. It was made by using a monte carlo method of generating random energies and then creating a histogram. The RMS value of each strain distribution is given by (3.2.11) and (3.2.12). In this case there are clearly two separate energy bands. When the entire valence band energy distribution is plotted the leading hydrostatic term $a\sqrt{3}\varepsilon_1$, causes the distribution to partially smear and two bands are not as distinct as shown in the dashed line of Figure 3.4. Finally, if the band gap distribution, as defined in the following section, is plotted the two peaks are completely smeared out and only a single peak is seen and shown in the dotted line of Figure 3.4.

3.3.2.2 Response of Bulk Silicon to Thermal Strains

In order to get an estimate of the variation of the band edges we will find the standard deviation of the band gap for a Δ conduction band minimum ($\Delta E_g = E_c - E_V = \Delta E_c - \Delta E_V$). First we redefine the energy shifts in terms of ε_1, ε_2, and ε_3 (3.3.5). These strains are orthogonal linear combinations of ε_{xx}, ε_{yy}, and ε_{zz}. Since ε_{xx}, ε_{yy}, and ε_{zz} are independent and equally distributed, ε_1, ε_2, and ε_3 are also independent and equally distributed. By construction, they even have the same distribution as ε_{xx}, ε_{yy}, and ε_{zz}. Redefining (3.3.1) and (3.3.2) using (3.3.5) gives the following conduction band energy shifts:

$$\Delta E_{c,av} = \left(\Xi_d + \frac{1}{3}\Xi_u\right)\sqrt{3}\varepsilon_1 \tag{3.3.6}$$

$$\Delta E_c^{100} - \Delta E_{c,av}^{100} = -\frac{\sqrt{6}}{3}\Xi_u^{\Delta}\left(-\frac{1}{2}\varepsilon_2 - \frac{\sqrt{3}}{2}\varepsilon_3\right)$$
$$\Delta E_c^{010} - \Delta E_{c,av}^{100} = -\frac{\sqrt{6}}{3}\Xi_u^{\Delta}\left(-\frac{1}{2}\varepsilon_2 + \frac{\sqrt{3}}{2}\varepsilon_3\right) \quad (3.3.7)$$
$$\Delta E_c^{001} - \Delta E_{c,av}^{100} = -\frac{\sqrt{6}}{3}\Xi_u^{\Delta}\varepsilon_2$$

The band gap will be defined by the CB minima distribution and the VB maxima distribution at any given time. In the conduction band, the multiple minima have the same distribution as a single minimum in the band tail and so we only need to consider one of the conduction band minima. However, the two valence band minima have different distributions and so we need to consider both of them. Consequently we get the following expression for the fluctuations in the band gap of silicon:

$$\Delta E_g = \left[\left(\Xi_d^{\Delta} + \frac{1}{3}\Xi_u^{\Delta}\right) - a\right]\sqrt{3}\varepsilon_1 + -\frac{\sqrt{6}}{3}\Xi_u^{\Delta}\varepsilon_2 \pm \sqrt{\left(\frac{\sqrt{6}}{2}b\varepsilon_2\right)^2 + \left(\frac{\sqrt{6}}{2}b\varepsilon_3\right)^2 + d^2\left[\varepsilon_{xy}^2 + \varepsilon_{yz}^2 + \varepsilon_{xz}^2\right]} \quad (3.3.8)$$

In calculating the standard deviation, $\sigma(\Delta E_g) = \sqrt{<\Delta E_g{}^2> - <\Delta E_g>^2}$, all of the terms can be found analytically as the mean of each strain is zero and the mean squared strains are given by (3.2.11) and (3.2.12). The mean squared values of ε_1, ε_2, and ε_3 will be the same as ε_{xx}. Equations (3.2.11) and (3.2.12) are only defined to within the factor β and so the standard deviation will be defined with respect to β as well. Nevertheless, this will still be useful for comparing to the following engineered cases. When accounting for both valence bands the mean of +/- the square root term is zero and the mean squared value follows from the mean squared values of the individual strains. Using deformation potentials from[36] and elastic constants from[35] we get $\sigma(\Delta E_g) = 0.229\sqrt{\beta}$ eV.

3.3.2.3 Engineering a Si-Ge Heterostructure [001] Device

By using bias strains or confinement it is possible to reduce the band gap fluctuations and

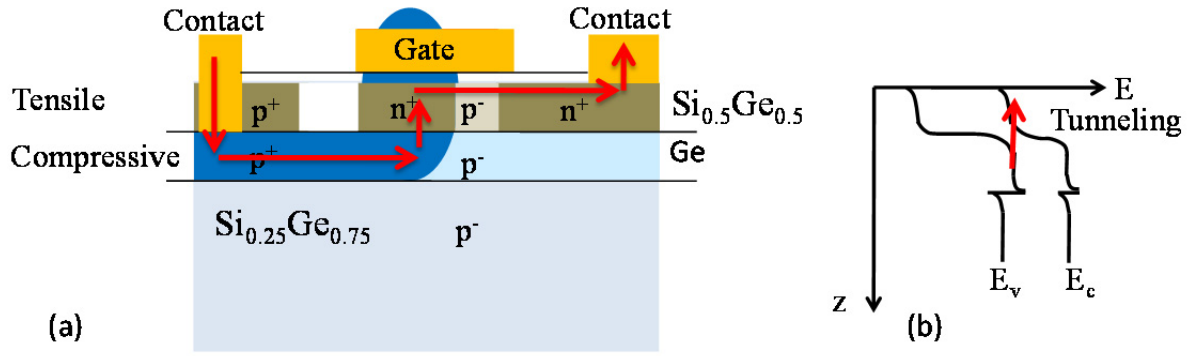


Figure 3.5: (a) A possible device structure using a phonon engineered heterostructure (b) the band diagram of the heterostructure

thus the subthreshold slope. Since there is an approximately uniform thermal strain across both the P and N sides of the tunneling junction for a given electron, the effective band gap will not change if somehow the conduction band on the N side and the valence band on the P side have the same energy shift in response to a strain.

Both the valence band and the conduction band [001] minima have the strain term ε_2. It is possible to get these terms to parallel each other with the correct biases. The [001] conduction band minima need to be lowered in energy with respect to the other minima. This means that a biaxial tensile strain needs to be applied to an Si N-region, possibly by growing the device on a Ge or Si-Ge substrate. Let $\varepsilon_2 = \varepsilon_2^0 + \varepsilon_2'$ where ε_2^0 is the grown in strain and ε_2' is the strain due to thermal vibrations. A biaxial tensile strain means $\varepsilon_2^0 > 0$. On the P-side of the device, strong confinement or a large ε_2^0 will suppress all of strain terms except for ε_2'. However, in order to have the conduction band parallel the valence band, the heavy hole ($J=3/2$, $m_J=3/2$) band must be raised above the light hole band. This means any bias strain must be compressive[24]. Thus the P-region could be germanium. A possible transistor structure based on this is shown in Figure 3.5. In this case the valence band edge energy will become:

$$E_v(\vec{k},\vec{\varepsilon}) = a\sqrt{3}\varepsilon_1 - \frac{k_z^2\hbar^2(\gamma_1 - 2\gamma_2)}{2m_0} + \frac{\sqrt{6}}{2}b(\varepsilon^0 + \varepsilon') \tag{3.3.9}$$

Combining the effects in the conduction band and valence band gives the following band gap fluctuation:

$$\Delta E_g = \left(\Xi_d^\Delta + \frac{1}{3}\Xi_u^\Delta\right)\sqrt{3}\varepsilon_1^{\,N} - a\sqrt{3}\varepsilon_1^{\,P} - \frac{\sqrt{6}}{3}\Xi_u^\Delta\varepsilon_2^{\,N} - \frac{\sqrt{6}}{2}b\varepsilon_2^{\,P} \tag{3.3.10}$$

where N stands for strains on the N-side and P for strains on the P side. We assume the strains are coherent throughout the N and P sides, but that the magnitude changes based on the crystal parameters. Since $\Xi_u^\Delta > 0$ and $b < 0$, the net effect is that the fluctuations are reduced. Therefore, we get $\sigma(\Delta E_g) = 0.132\sqrt{\beta}$ eV. Consequently, growing a device with biaxial tensile strain in the N-region and strong confinement or compressive strain in the P-region, results in a 42%

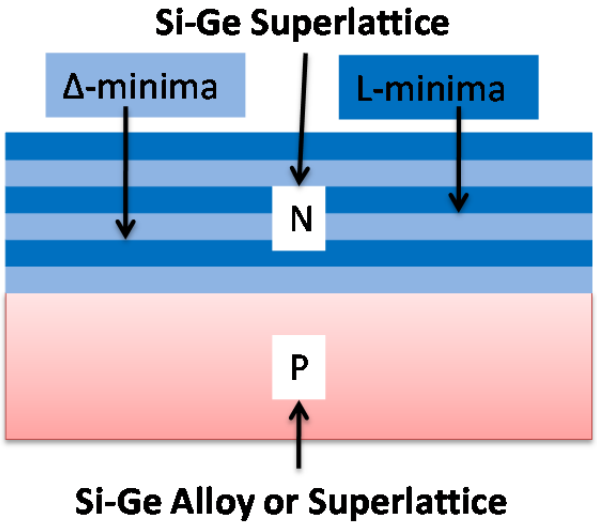


Figure 3.6: Tunneling junction with Si-Ge superlattice

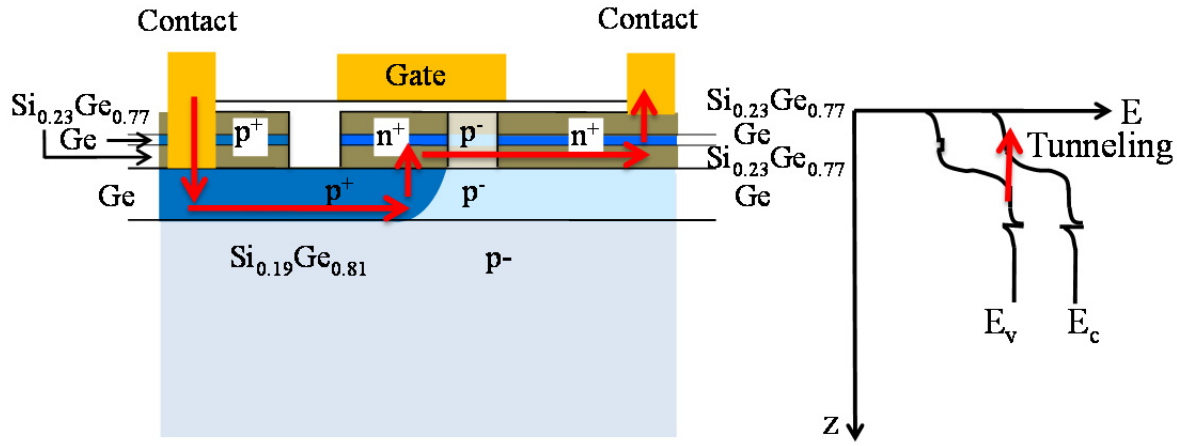


Figure 3.7: (a) TFET based on a SiGe superlattice (b) Band diagram along the tunneling junction

reduction in the effect of long wavelength acoustic phonons. Thus a significant improvement can be achieved in simple silicon (n-side) germanium (p-side) heterostructure.

3.3.2.4 Engineering a Si-Ge Superlattice [001] Device

If we have more control over deformation potentials it is be possible to reduce the variations in the band gap even more. Introducing a second material, germanium, in an alloy or superlattice gives us this control. One especially interesting feature of germanium is that the hydrostatic band gap deformation potential, $\left(\Xi_d^L + 1/3\,\Xi_u^L\right) - a$, is negative while in silicon it is positive.

However, the deformation potentials in the conduction band are closely related to what type of minima are the lowest energy minima (i.e. Δ or L) and so deformation potentials will not scale linearly in an alloy. Consequently a short period superlattice is necessary in order to have more of a linear interpolation between the two materials. In order to have both the Δ and L states mix, they must both be degenerate in energy[39]. Thus the superlattice will consist of Si-Ge alloys, with a Si-like part that corresponds to the Δ minima and a Ge-like part that corresponds to the L minima. For ease of fabrication we consider a strain relaxed superlattice with the L part consisting of pure germanium. We then need to match the strained conduction band energy of the Δ part to that of the L/Ge part. This depends on the value of the strain which depends on the optimized superlattice composition as described later. Calculating the energies using the model solid theory [36], with band gap bowing given by[40], and elastic constants from [35] for a superlattice that is 80% Si-like and 20% Ge, results in the si-like region being composed of 23% Si and 77% Ge. Since the valance band in both materials is of the same type, either an alloy or a short period superlattice can be used on the P side. A generalized schematic is shown in Figure 3.6. Figure 3.7 shows a TFET structure built using the superlattice.

As a simplifying approximation we will neglect phonon scattering off the superlattice and assume that Δ and L regions have linearly interpolated bulk properties. We also assume that the phonon modes are coherent through both the Δ and L regions, but that the amplitude changes as the elastic constants change. Furthermore, we will use the same bias strain methods used in the previous section. Thus the Δ parts of the N region must be under tensile strain and the P region

must be under a compressive strain or strong confinement. We get the following expression for the conduction band energy:

$$\begin{aligned}\Delta E_c &= x\left[\left(\Xi_d^{\Delta}+\frac{1}{3}\Xi_u^{\Delta}\right)\sqrt{3}\varepsilon_1^{\Delta}-\frac{\sqrt{6}}{3}\Xi_{u,}^{\Delta}\varepsilon_2^{\Delta}\right]\\&+(1-x)\left(\Xi_d^{L}+\frac{1}{3}\Xi_u^{L}\right)\sqrt{3}\varepsilon_1^{L}\\&+(1-x)\frac{2}{3}\Xi_u^{L}\left(\varepsilon_{xy}^{L}+\varepsilon_{yz}^{L}+\varepsilon_{xz}^{L}\right)\end{aligned} \tag{3.3.11}$$

where x is the fraction of si-like/ Δ material on N side. The valence band energy shift is:

$$\begin{aligned}\Delta E_v &= y\left(a_{si}\sqrt{3}\varepsilon_1^{si}+\frac{\sqrt{6}}{2}b_{si}\varepsilon_2^{si}\right)\\&+(1-y)\left(a_{ge}\sqrt{3}\varepsilon_1^{ge}+\frac{\sqrt{6}}{2}b_{ge}\varepsilon_2^{ge}\right)\end{aligned} \tag{3.3.12}$$

where y is the fraction of silicon on the P side. Calculating the standard deviation of $\Delta E_g = \Delta E_c - \Delta E_V$ using the methods outlined in the previous sections and minimizing it with respect to x and y (while also calculating the strains and deformation potentials in the superlattice) gives the following results. We get $\sigma(\Delta E_g) = 0.0856\sqrt{\beta}$ eV with 80% si-like/ Δ material in the N side and 34% Si on the P-side. The Si-like/Δ material is composed of 23% Si / 77% Ge and the L material is pure Ge. This results in a 63% improvement compared to bulk silicon. This is shown in Figure 3.7. Interestingly, the composition of the P region has a small effect on the standard deviation. Arbitrarily changing the composition changes $\sigma(\Delta E_g)$ by no more than 5%. This is because the different types of CB minima in the N region have significantly different deformation potentials, while the valence band maxima are of the same type. Nevertheless a germanium rich P region is necessary in order to correctly split the valance band maxima.

3.4 CONCLUSIONS

As shown in the previous sections, it may be possible to get roughly a 60% reduction in the band gap fluctuations due to long wavelength acoustic phonons and thus a corresponding reduction in the subthreshold slope. This could achieved by growing a device under tensile strain with a short period superlattice of roughly 80% si-like material and 20% Ge in the N side. The si-like material can be composed of 23% Si/ 77% Ge. The composition of the p side can engineered to improve other device properties, so long as it is compressively strained or confined. The exact compositions will have to be determined experimentally as there is still a large variation in the values of the deformation potentials in the literature. However, the experimentally realized gains may be smaller as a number of approximations were made in this derivation. In particular the assumption of a uniform strain over the entire coherence volume of an electron may not be entirely true, and phonon scattering and superlattice effects have not been fully accounted for. Despite these limitations, this work shows that there is a strong possibility of improving the subthreshold slope using just silicon and germanium.

3.5 APPENDIX A- CALCULATING RMS STRAINS USING PHONON MODES

In order to find the total RMS strain ε_{ij}, the strain contribution from each phonon needs to be added. This is done by first discretizing k-space by using periodic boundary conditions, i.e. $k_i = \pm 2n\pi / L_i$ where L_i is the length of the i'th dimension of the crystal and n=0,1,2,… Then a linear acoustic phonon dispersion relationship is assumed and so the equations of motion for sound in a solid can be solved in order to give the phonon modes (3.2.1). The equations of motion are:

$$\rho \frac{\partial^2 \delta R_x}{\partial t^2} = C_{11} \frac{\partial^2 \delta R_x}{\partial x^2} + C_{44}\left(\frac{\partial^2 \delta R_x}{\partial y^2} + \frac{\partial^2 \delta R_x}{\partial z^2} \right) + (C_{12} + C_{44})\left(\frac{\partial^2 \delta R_y}{\partial x \partial y} + \frac{\partial^2 \delta R_z}{\partial x \partial z} \right) \tag{3.5.1}$$

$$\rho \frac{\partial^2 \delta R_y}{\partial t^2} = C_{11} \frac{\partial^2 \delta R_y}{\partial y^2} + C_{44}\left(\frac{\partial^2 \delta R_y}{\partial x^2} + \frac{\partial^2 \delta R_y}{\partial z^2} \right) + (C_{12} + C_{44})\left(\frac{\partial^2 \delta R_x}{\partial x \partial y} + \frac{\partial^2 \delta R_z}{\partial y \partial z} \right) \tag{3.5.2}$$

$$\rho \frac{\partial^2 \delta R_z}{\partial t^2} = C_{11} \frac{\partial^2 \delta R_z}{\partial z^2} + C_{44}\left(\frac{\partial^2 \delta R_z}{\partial x^2} + \frac{\partial^2 \delta R_z}{\partial y^2} \right) + (C_{12} + C_{44})\left(\frac{\partial^2 \delta R_x}{\partial x \partial z} + \frac{\partial^2 \delta R_y}{\partial y \partial z} \right) \tag{3.5.3}$$

Solving for $\delta\vec{R} = R_x\hat{x} + R_y\hat{y} + R_z\hat{z} = (A_x\hat{x} + A_y\hat{y} + A_z\hat{z})\cos(k_x x + k_y y + k_z z - \omega t)$ at each point in k-space simplifies to a simple eigenvalue problem as the displacements are sinusoidal. The result is three eigenvectors that correspond to the longitudinal and transverse phonon modes. This gives the relative values of A_x, A_y, and A_z but not the overall magnitude, |A|. The magnitude can be found by using the fact that each phonon mode has $k_bT/2$ Joules of energy. The total energy is given by (3.2.4). However this equation is in terms of displacements and not strains and so the strains need to be found by plugging the displacement $\delta\vec{R}$ into (3.2.3). Thus the |A| is known and the strains due to each phonon mode $\varepsilon_{ij}(k,s)$ are known. Finally the total strain should be found by summing over all points in the 1st Brillouin Zone (BZ). For instance, $< \varepsilon_{xx}{}^2 >= \sum_{k \text{ in 1st BZ}} \sum_{s=1,2,3} < \varepsilon_{xx}(k,s)^2 >$ where s represents the three modes per point in k-space.

3.6 APPENDIX B- TABLE OF DEFORMATION POTENTIALS

Parameter name	Description	Si [36, 41]	Ge[36, 41]
a	Average valence band shift	2.46	1.24
b	Valence band splitting (100) strain	-2.35	-2.55
d	Valence band splitting (111) strain	-5.32	-5.50
Ξ_d^{Δ}	Conduction band – dilatation deformation potential - Δ minimum	1.13	-0.59
Ξ_u^{Δ}	Conduction band – uniaxial deformation potential - Δ minimum	9.16	9.42
$\Xi_d^{\Delta}+\frac{1}{3}\Xi_u^{\Delta}$	Conduction band – hydrostatic deformation potential - Δ minimum	4.18	2.55
$\Xi_d^{\Delta}+\frac{1}{3}\Xi_u^{\Delta}-a$	Band gap deformation potential - Δ minimum	1.72	1.31
Ξ_d^{L}	Conduction band – dilatation deformation potential - L minimum	-6.04	-6.58
Ξ_u^{L}	Conduction band – uniaxial deformation potential - L minimum	16.14	15.13
$\Xi_d^{L}+\frac{1}{3}\Xi_u^{L}$	Conduction band – hydrostatic deformation potential - L minimum	-0.66	-1.54
$\Xi_d^{L}+\frac{1}{3}\Xi_u^{L}-a$	Band gap deformation potential - L minimum	-3.12	-2.78

All values are in eV

Chapter 4: Modeling and Experimentally Determining the Band Edge Steepness

4.1 INTRODUCTION

In order to properly design a tunneling junction we need to know just how steep the band edges are. As mentioned in Section 3.1 it is possible to extrapolate the band edge steepness from the steepness optical absorption, or the Urbach tail. The optical absorption is proportional to the joint density of states and so if the optical absorption falls of exponentially, the density of states should follow the same pattern. Nevertheless, since an absorption measurement is inherently an

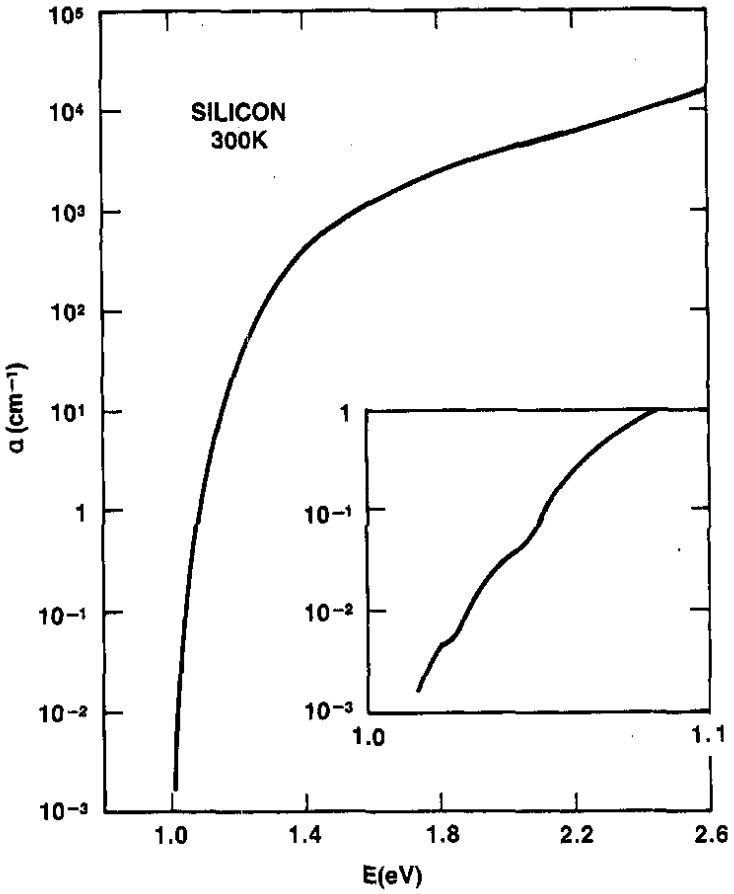


Fig. 3. Optical absorption coefficient of silicon at 300 K in the vicinity of the band edge [12].

Figure 4.1: Optical absorption coefficient of silicon at 300K in the vicinity of the band edge (from Tiedje 1984)

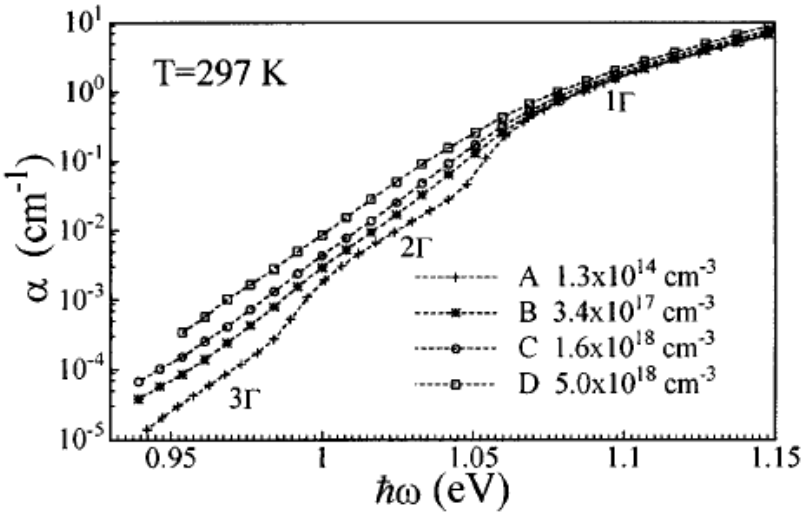


Figure 4.2: The absorption curves of Silicon at different doping levels (from Duab 1996)

optical process, it is possible that the physics of the tunneling process could be different. Consequently, it would be extremely useful to have an electronic way to measure the steepness of the band edges. To some extent this can be done by correctly interpreting the I-V characteristics of a backward diode.

In this chapter we will first go through some optical absorption measurements showing the limitations of different types of materials. Then we will show how to interpret backward diode measurements to extract a rough measure of the band edge steepness and how to model the band tail effects on the I-V curve. Finally we will compare a number of different engineered backward diodes to see what sort of band edge steepness we can expect.

4.2 MEASURING THE DENSITY OF STATES THROUGH OPTICAL ABSORPTION

In bulk undoped silicon the optical absorption silicon falls off at a rate of 27 mV/decade as seen from Figure 4.1 [25, 26]. Consequently, we can expect to see a similar steepness for the density of states turn on in a tunneling junction. The next thing we can ask is: what is the steepness for a doped semiconductor? Typically free carrier absorption will hide the band tails and. In order to avoid this, a type of photoluminescence experiment need to be done [42]. The resulting absorption measurements are shown in Figure 4.2. As seen in the figure varying the doping from 1e14 to 5e18 only seems to change the steepness of the absorption from 27 mV/decade to 30 mV/decade.

This seems to indicate that moderate doping will not have a significant effect on the steepness density of states. Unfortunately, the same results may not be true in tunneling junction. The theory modeling the doping induced band tails [43-47] is extremely sensitive to the electrostatic screening length. This means that having a lot of free carriers can screen out many of the potential fluctuations and reduce the impact of the band tails. Unfortunately, in a tunneling junction there are very few carriers in the depletion region. This means that a tunneling switch will not get the benefit of the electrostatic screening and so the actual band tails

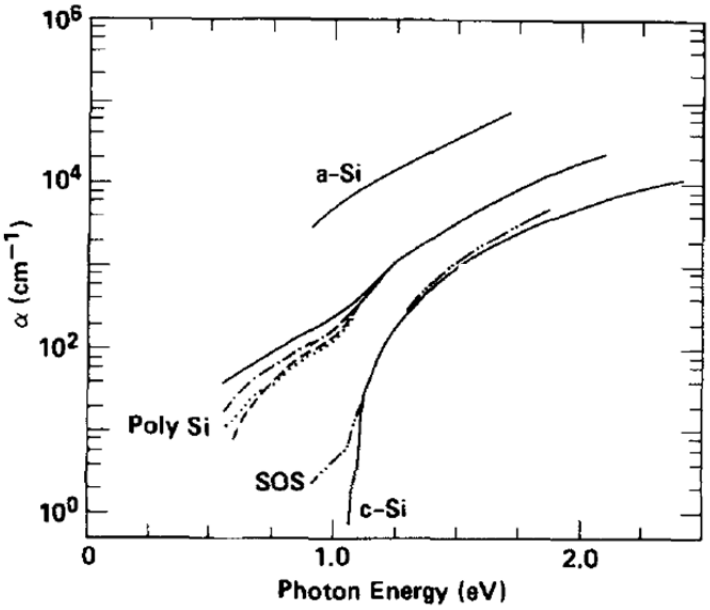


Figure 4.3: (from Jackson 1983) Absorption vs photon energy for different types of silicon. Crystalline silicon, silicon on sapphire (SOS), polycrystalline silicon and amorphous silicon are plotted. The different polycrystalline lines correspond to different intervals of hydrogen exposure varying from unhydrogenated (solid) to 120 min (dashed)

will be much worse. This can be seen when we analyze the diode measurements in the following sections. In order to see the correct band tails in an optical measurement we would need to measure the absorption of a compensated semiconductor that has an equal number of N and P type dopants. This will eliminate the free carriers that screen out the potential fluctuations. Even with the electrostatic screening, the models predict that the band tails will get significantly worse once the doping is in the 10^{19} to the 10^{20} range[43-47].

In order to see just how bad the band tails can get we can look at the optical absorption of undoped polycrystalline silicon [48]. The absorption is plotted in Figure 4.3. The absorption coefficient falls off with a slope of around 200 mV/decade. That is completely unacceptable for a tunneling switch. The reason for the poor slope can be seen from Figure 4.4. There is a broad density of states near the band edge that results from the poor crystal structure.

4.3 USING BACKWARD DIODES TO MEASURE STEEPNESS

When designing a TFET we are interested in how the gate voltage changes the conductance of the channel at a fixed source drain bias. Since the conductance is the relevant measure, we should look at how the conductance of a backward diode changes with bias. A backward diode cannot achieve the same level of electrostatic control over the tunneling junction as a transistor and so it is very difficult to achieve a steep response using barrier width modulation. However, almost all the voltage applied to a PN junction is dropped in the depletion region and so the change in band alignment directly corresponds to the applied voltage. This

means that the change in the conductance will be predominantly limited by the density of states turn on.

The benefit of looking at the conductance can be seen by looking at a modified equation for the tunneling current:

$$j_t \propto \int \partial E * T * (f_C - f_V) * D_{BandEdge}(E) \tag{4.3.1}$$

In order to account for the band edge density of states we have added an extra band edge density of states term. We will refine this model further in the following section. The key difference between a transistor and a diode is the Fermi function term, $(f_C - f_V)$. In a transistor it is fixed by the source drain bias and is a constant, but it varies with bias in a diode. In order to see the transistor response in a diode we need to divide out the effect of the Fermi functions. At small biases it is easy as we can Taylor expand the Fermi functions:

$$(f_C - f_V) \approx \left(\frac{1}{2} - \frac{E - F_C}{4K_B T}\right) - \left(\frac{1}{2} - \frac{E - F_V}{4K_B T}\right) = \frac{F_C - F_V}{4k_b T} = \frac{qV}{4k_b T} \tag{4.3.2}$$

This means that simply dividing by the voltage and plotting the conductance will give the transistor response!

At larger biases the method will only be approximately correct as the Fermi functions will vary with energy. Nevertheless the Fermi functions still introduce a proportionality to the applied voltage as:

$$\int (f_C - f_V) dE = qV \tag{4.3.3}$$

This means that dividing by the applied voltage is still helps to remove the effect of the Fermi functions at higher biases. Of course this method still does not separate the effects of barrier modulation and band edge steepness. Nevertheless, useful information can still be gathered by correctly interpreting the data as seen in the next section.

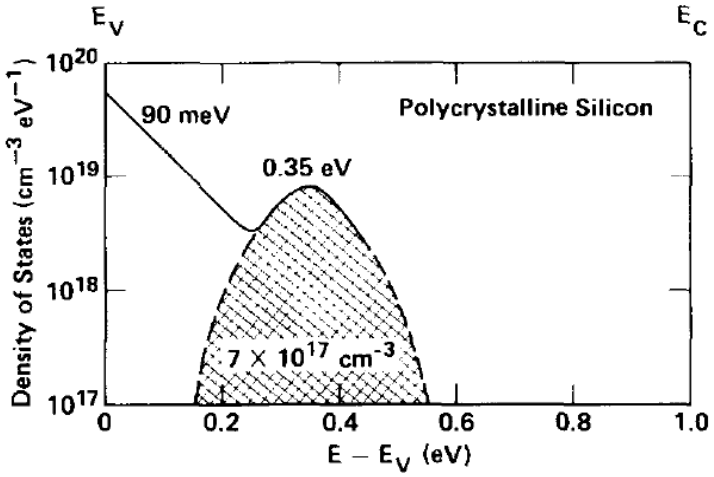


Figure 4.4: (from Jackson 1983) Features of the grain boundary density of states derived from optical absorption and ESR measurements for polysilicon.

4.3.1 Measuring the Steepness of Silicon Backward Diodes

Now we can apply this method to some backward diodes build by IBM [49]. Dr. Solomon shared the raw data with us so that we could analyze the diodes. The structure is a simple implanted diode shown in Figure 4.5. In order to have an abrupt heavily doped junction the diode was heavily implanted to near 1e20 and is heavily compensated. This means that there are an excessive number of dopant atoms present that will create a large impurity band that has a gradual density of states tail extending into the band gap. This is reflected in the electrical measurements.

Figure 4.6(a) shows the I-V curve. As expected it shows a little bit of Esaki behavior that is more pronounced at low temperatures. The I-V curve appears to be very steep at zero bias, but that is only because the current crosses zero at zero bias. Figure 4.6(b) shows the plot of the conductance which has two very distinct regions as labeled in the figure. By plotting the conductance we are able to use the data near the zero crossing and in the Esaki part of the curve. Figure 4.6(c) shows the slope of the conductance in mV/decade. The dotted line at the bottom of the plot indicates the desired slope of 60 mV/decade. Since the conductance is a reflection of the performance one could hope to achieve in a transistor, the slope of the conductance indicates the subthreshold slope that a transistor could achieve using the density of states switching mechanism. A transistor can have better performance using barrier width modulation, but the density of states switching will have the same performance in a diode or a transistor as only the band alignment matters.

The diode's initial turn-on is dominated by the density of states turn on and is reflected by the steeper region in Figure 4.6(b) and (c). At the higher current levels, the current is limited by the tunneling probability and so it is the barrier width modulation that controls the shape of the I-V curve. This is also indicated in Figure 4.6(b) and (c). Unfortunately the minimum conductance slope is around 120 mV/decade. This means that the heavy doping has caused severe band tails that limit the steepness of the band edges.

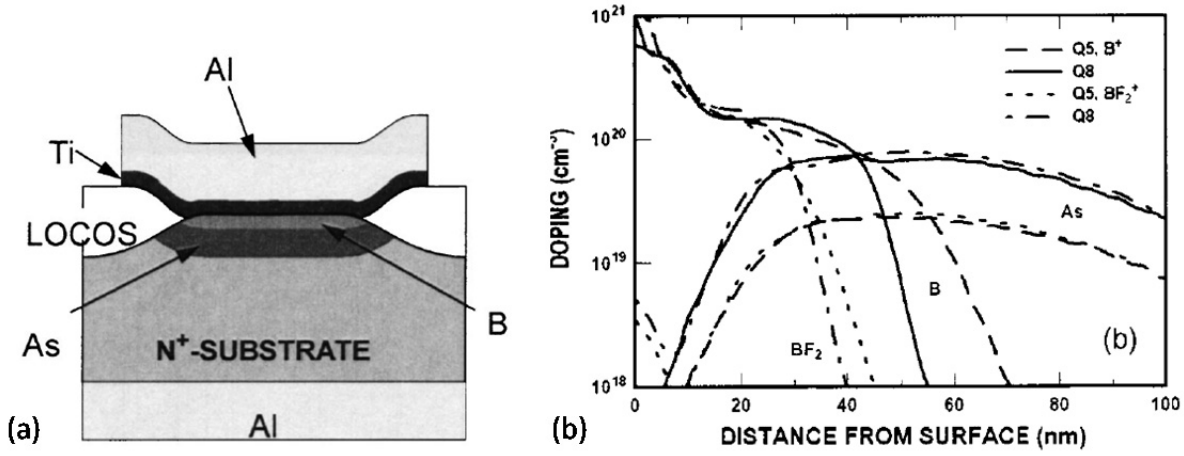


Figure 4.5: (from Solomon 2004) Backward diodes built in silicon. (a) The structure is a simple implanted PN junction. (b) The doping levels are near 1e20 and the diode is heavily compensated

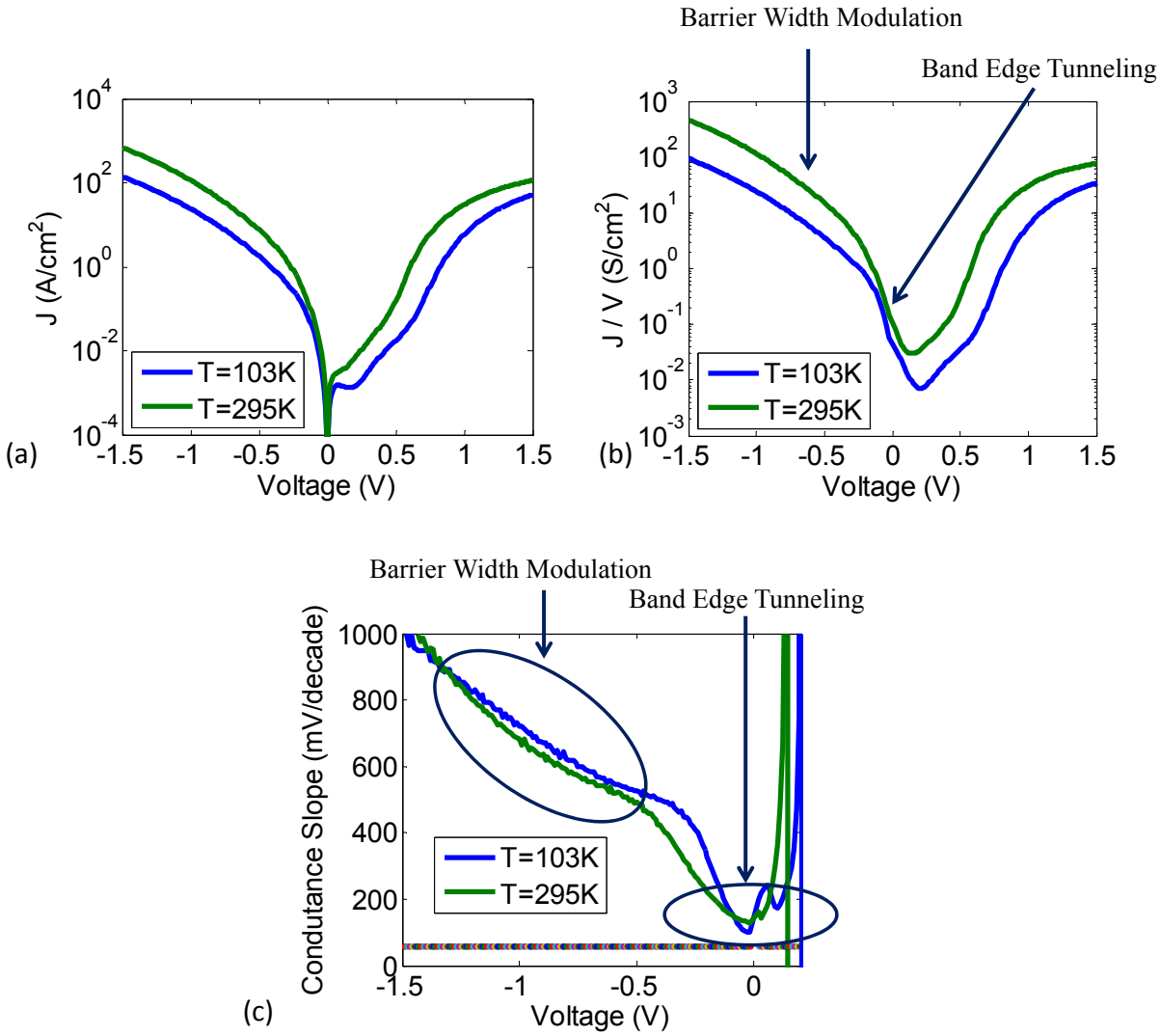


Figure 4.6: (a) The I-V characteristics for a silicon backward diode measured at two temperatures. (b) The conductance for the backward diode. There are two regions in the conductance curve corresponding to band edge tunneling and barrier width modulation. (c) The slope of the conductance in mV/decade is plotted. The two different regions are more evident in the plot of the conductance slope.

4.3.2 A Refined Band Tail Model

Now that we have I-V curves we can create a simple model to verify our interpretations:

$$J = J_0 \int \partial E \times T \times (f_C - f_V) \times D_n(E) \times D_p(E) \tag{4.3.4}$$

In this model we will ignore the effects of transverse momentum and simply defined a constant prefactor to account most phonon related effects. In this model we only keep a few critical terms. The voltage dependent tunneling probability is:

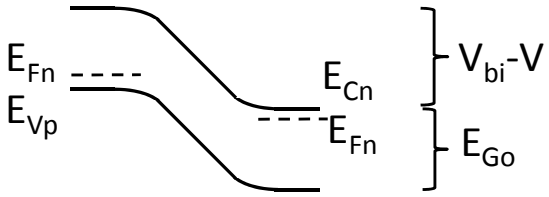


Figure 4.7: The various energy levels used in the tunneling model are depicted here

$$T(V_{SD}) = \exp\left(\frac{-\alpha}{(V_{bi}-V)^{1/2}}\right) \tag{4.3.5}$$

Here α is a fitting parameter that allows us to fit the barrier modulation tunneling. The equation is derived from the Kane tunneling probability given in Eqn (2.2.2.) [22]:

$$T = \exp\left(\frac{-\pi(m^*)^{1/2}E_G^{3/2}}{2\sqrt{2}\hbar q\vec{E}}\right) \tag{4.3.6}$$

In this case we are interested in the bias dependence which is given in the electric field term which is the peak electric field in a PN Junction:

$$\vec{E} = \left(\frac{2q}{\varepsilon}\left(\frac{N_D N_A}{N_D + N_A}\right)(V_{bi}-V)\right)^{1/2} \tag{4.3.7}$$

Since many of the parameters are uncertain we abstracted everything except for the key $(V_{bi}-V)$ dependence as a single fitting parameter. V_{bi} is the built in voltage across the junction.

Next, the Fermi functions are given by the standard equations:

$$f_c = \frac{1}{1+e^{(E-E_{fn})/k_bT}} \tag{4.3.8}$$

$$f_v = \frac{1}{1+e^{(E-E_{fp})/k_bT}} \tag{4.3.9}$$

The Fermi levels are depicted in Figure 4.7. Finally we need to define the band tail density of states:

$$D_p(E) = \begin{cases} 1, & E < E_{Vp} \\ e^{-(E-E_{Vp})/qV_O}, & E > E_{Vp} \end{cases} \tag{4.3.10}$$

$$D_n(E) = \begin{cases} 1, & E > E_{Cn} \\ e^{+(E-E_{Cn})/qV_O}, & E \le E_{Cn} \end{cases} \tag{4.3.11}$$

Here E_{Vp} is the valence band edge on the p side and E_{Cn} is the conduction band edge on the n side as depicted in Figure 4.7. In defining these band edge functions we assumed that the density of states fell of exponentially from the band edge with some slope V_O. This slope gives the steepness of the density of states band tail and ideally would be less than K_bT (or 60 mV/decade) for a good tunnel junction.

4.3.2.1 Applying the Band Tail Model to Silicon Diodes

Now that we have a model for the band tails we can apply it to the Esaki diodes that we previously analyzed. In order to fit the IBM diode data we used the following values for the constants in the model:

V_{bi} =0.91 V

E_{Go} =1.12 eV

V_o =0.126 V

J_0 =3.84e7 A/cm^2

α =0.843 V$^{1/2}$

As seen in Figure 4.8, the analytic model and the experimental data fit very well. Of course, this is mostly due to the fitting parameters. Nevertheless, this gives us some confidence in our interpretation of the different regions of the I-V curve. One thing to note is that the exponential slope required to fit the date, V_0, is 0.126 V. This corresponds to a slope of 0.126*ln(10)=290 mV/decade. This is far worse than the minimum slope of 120 mV/decade that was extracted directly from the experimental data. The reason for this is that the model includes both the effects of barrier modulation and band edge steepness. When the two effects are multiplied together, the resulting slope can be slightly better than either one taken individually. Nevertheless, analyzing the conductance slope can still provide a lower limit for the density of states steepness.

Since both the band edge steepness and the barrier thickness modulation play a key role in determining the overall steepness, models like this one need to be developed to more accurately account for both effects. Doing so would allow us to explore the sort of optimizations between barrier width modulation and band edge steepness needed at the end of Chapter 2.

4.3.3 Using the Backward Diode Figure of Merit, γ

So far, the silicon diodes that were analyzed were not optimized for being good backward

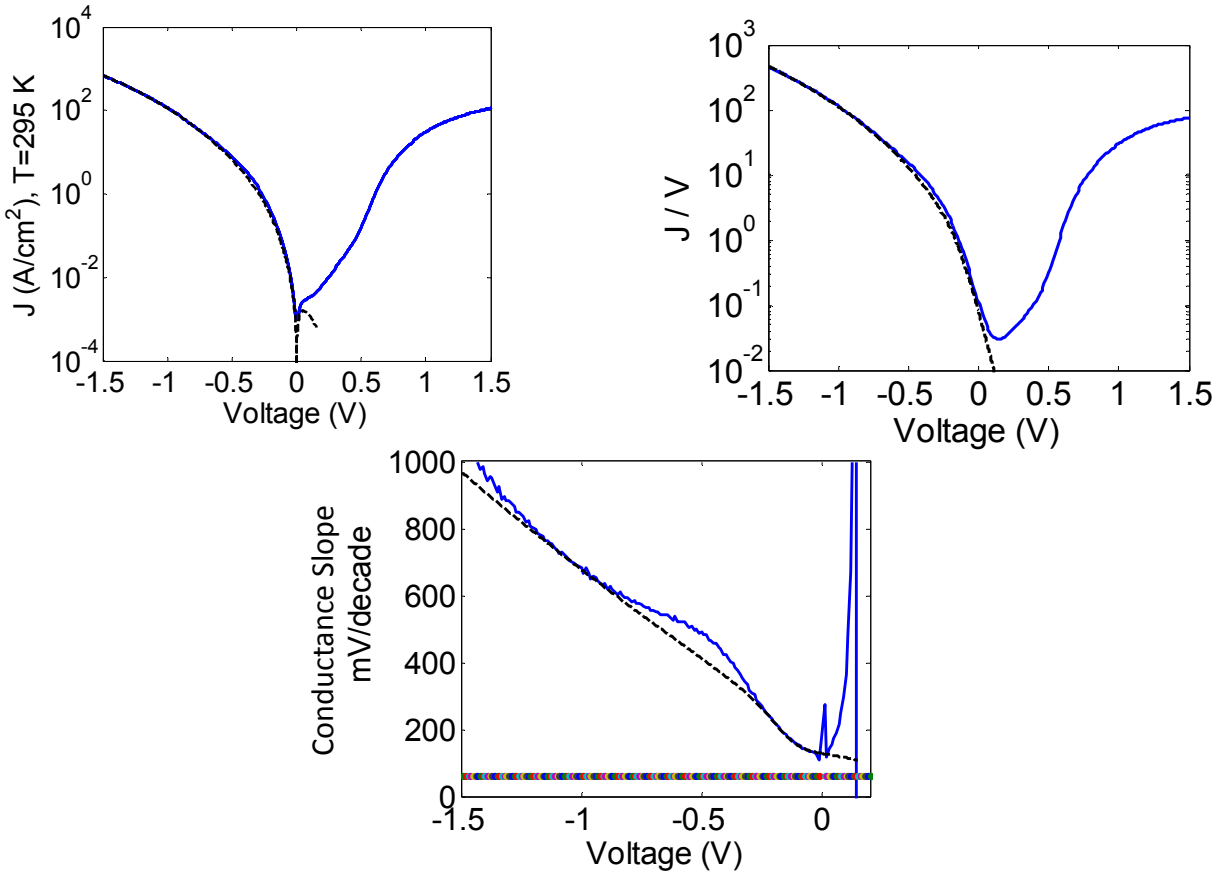


Figure 4.8: Comparison between the fitted analytic model and measured data for a silicon backward diode. (a) The reverse bias I-V curves match very well. The forward bias current was not modeled. (b) The model also fits the conductance

diodes with a highly non-linear characteristic. The figure of merit for backward diodes is called γ and is defined as:

$$\gamma = \frac{d^2 I / dV^2}{dI / dV} \tag{4.3.12}$$

In order to relate this to the steepness of the tunneling junction we need a model for the tunneling current. Near zero bias the current has two key dependencies on the voltage. First the difference between the Fermi functions $(f_C - f_V)$ goes to zero and can be Taylor expanded as done in Eqn (4.3.2). Second, the tunneling current is exponentially increasing with increasing reverse bias. Consequently we have

$$I \propto (f_C - f_V) \times e^{-V/V_O} \propto V \times e^{-V/V_O}$$

Here V_0 is the exponential steepness that we are interested in. Plugging this into (4.3.12) gives:

$$\gamma = \frac{d^2 I / dV^2}{dI / dV} = -2 / V_O$$

We can then convert this to a conductance slope in mV/decade:

$$\begin{aligned} \text{Conductance Slope(mV/decade)} &= \ln(10) \times V_O \\ &= \ln(10) \times 2 / \gamma \end{aligned}$$

Using this, we see that we need γ>80 in order to have a conductance slope less than 60 mV/decade.

In order to check this method we can plot the slope of germanium a backward diode [15]and compare the extracted steepness to the steepness derived from γ. The I-V curve in the paper is digitized and replotted in Figure 4.9. The conductance and conductance slope is also plotted. The average conductance slope is 92 mV/decade. This is identical to the slope computed from γ: $2/\gamma \times \ln(10)$ = 92 mV/decade.

Now, looking through the backward diode literature for the best γ ever reported we find a γ around 70 from a 1967 paper by Karlovsky [16]. Unfortunately, this is still not greater than 80. Does this mean that the band edges cannot be sharper than 60 mV/decade? Most likely this is the case for current backward diodes that require heavily doped junctions to get a turn on near zero bias.

4.3.4 Comparing Backward Diodes

In the next few pages we plot the current, conductance and slope for a variety of different diodes to see how they compare. In Figure 4.10 we analyze the best (highest γ) InAs / AlGaSb backward diode reported to date[19]. Since the width of the tunneling barrier is fixed by the AlSb, the slope of this device is determined entirely by the density of states turn on. Unfortunately the slope is still worse than the 60 mV/decade limit. Figure 4.11 shows an MBE

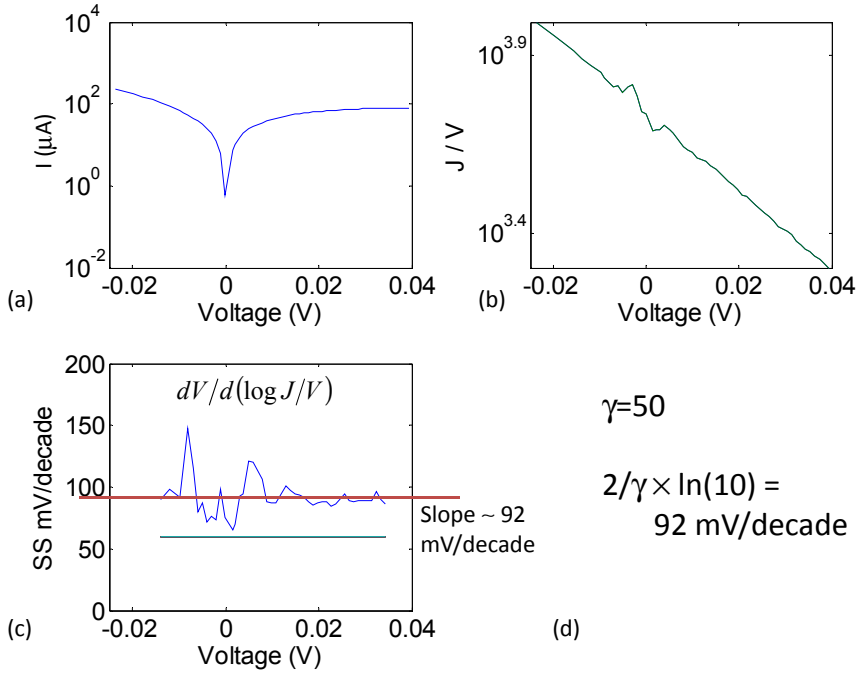


Figure 4.9: (after Karlovsky 1961) (a) The I-V characteristics near zero bias for a germanium backward diode are plotted. (b) The conductance falls off exponentially with voltage.

grown silicon Esaki diode. Figure 4.13 is an $In_{0.53}Ga_{0.47}As$ PIN Diode and Figure 4.12 is an InAs nanowire grown on silicon. None of the diodes show sub 60 mV/decade slopes.

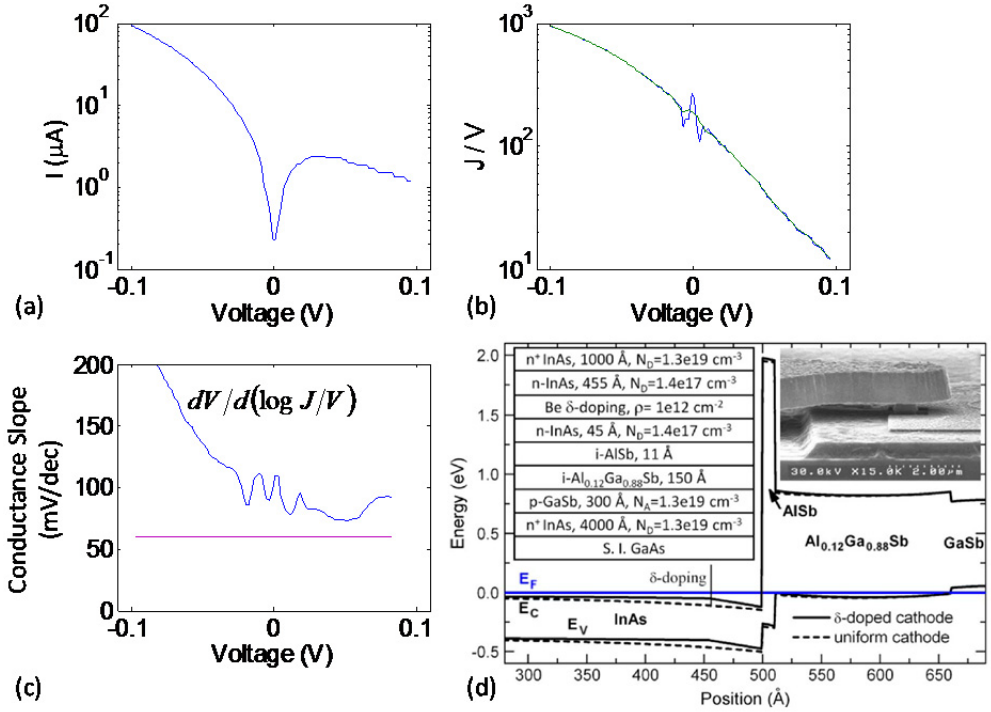


Figure 4.10: The (a) current (b) conductance and (c) conductance slope for an InAs/AlGaSb backward diode is plotted. (d) The device structure is shown (from Zhang 2011)

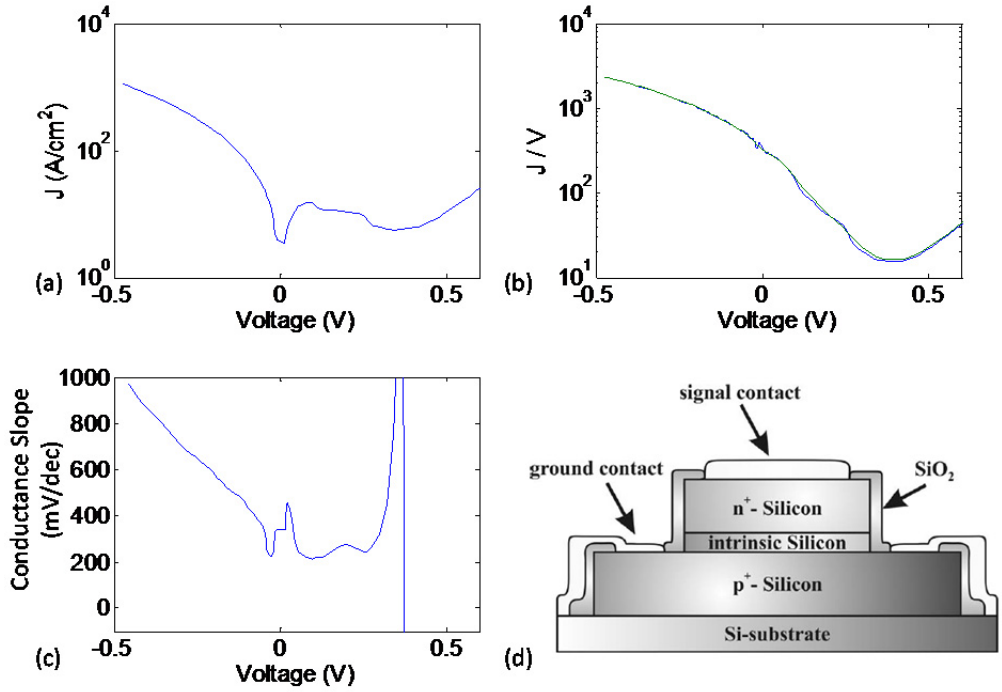


Figure 4.11: The (a) current (b) conductance and (c) conductance slope for an MBE grown silicon Esaki diode is plotted. (d) The device structure is shown (from Oehme 2009)

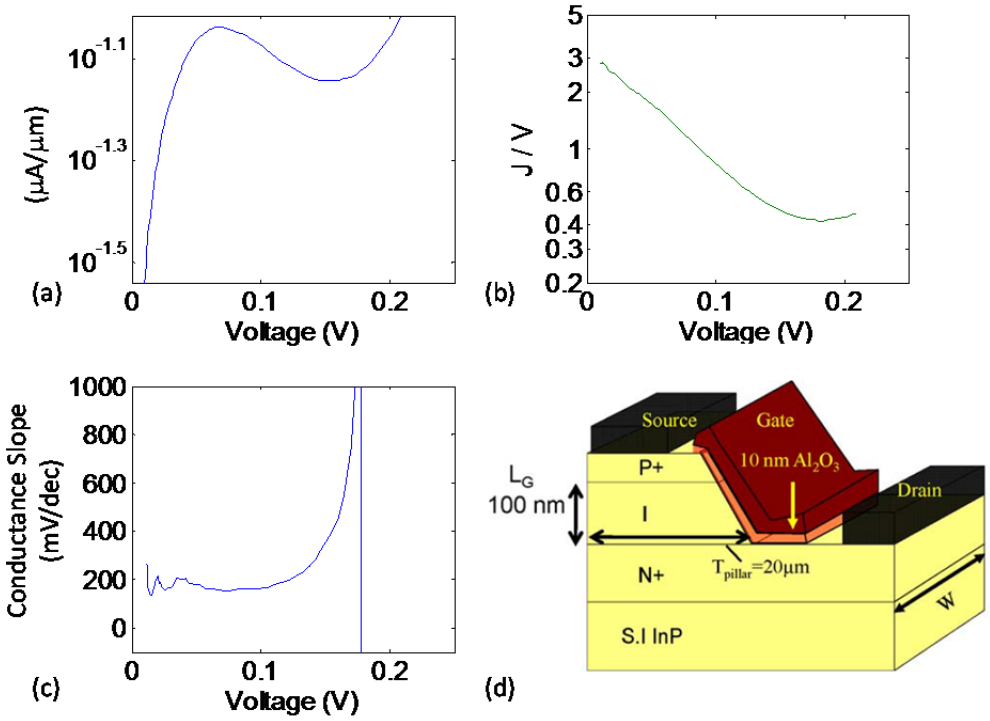


Figure 4.13 The (a) current (b) conductance and (c) conductance slope for an $In_{0.53}Ga_{0.47}As$ PIN Diode is plotted. (d) The device structure is shown (from Tomioka 2011)

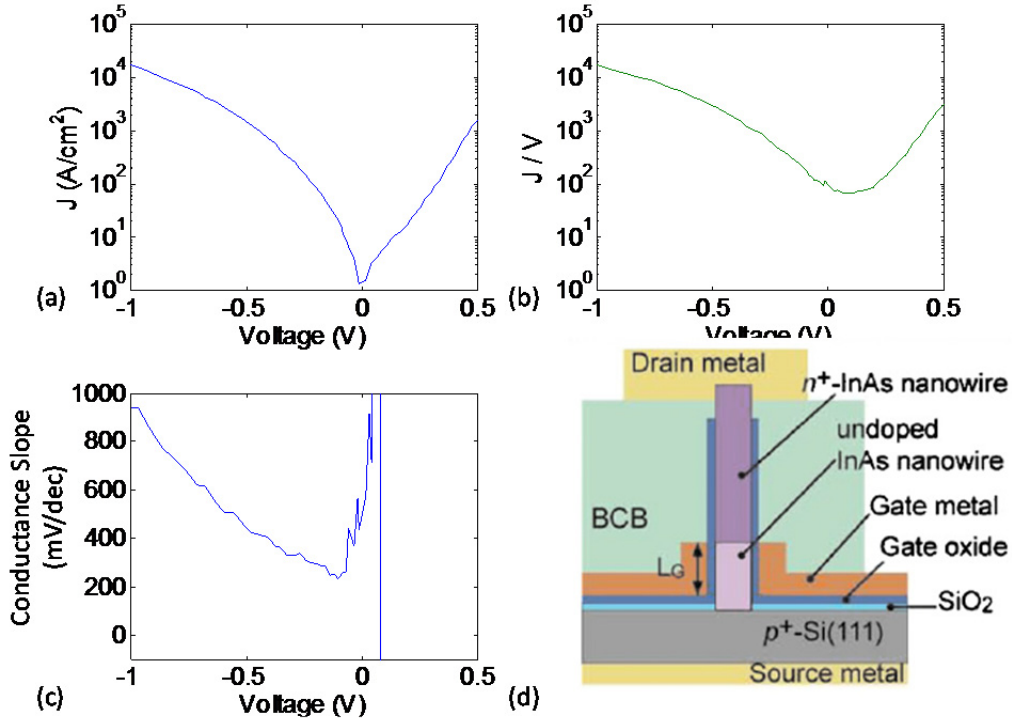


Figure 4.12: The (a) current (b) conductance and (c) conductance slope for an InAs nanowire on silicon is plotted. (d) The device structure is shown (from Tomioka 2011)

4.4 CONCLUSION

It is critical to preserve the quality of the semiconductor to ensure a sharp band edge. This can be seen from both the optical and electrical measurements. If the crystalline quality is ruined through either a polycrystalline semiconductor or through heavy doping, a long tail of states extending into the band gap will form.

Furthermore, additional models like Eqn (4.3.4) that account for both the band tails and barrier width modulation need to be developed in order to fully analyze the performance of a tunneling junction. If a model like Eqn (4.3.4) is implemented in TCAD tools, we should be able to get significantly better and more useful simulation results.

Finally, we still need more ways to measure the band edge density of states. A method such as scanning tunneling spectroscopy might be useful for measuring the local density of states. It might also be possible to interpret the smearing of a C-V measurement to give the density of states falling into band gap. We can infer information about the density of states, but neither optical nor the electrical measurements are unambiguous.

Chapter 5: Trap Assisted Tunneling and Other Limitations on Tunneling Switches

5.1 Trap Assisted Tunneling

Trap assisted tunneling has been a major limitation of many experimental TFET results [50] and can often explain an anomalous temperature dependence in the threshold voltage of TFETs. Consequently, we present a model for trap assisted tunneling and use it to explain the anomalous temperature dependence of silicon pocket TFETs. The different types of current in a reverse biased PN junction are shown in Figure 5.1. The standard thermoionic current is shown in orange. The direct tunneling current we're interested in is shown in blue. There are also two leakage currents that pass through trap states in the band gap. The first is the result of the standard Shockley-Reed-Hall (SRH) generation where an electron is thermally excited from the valence band to a trap and then it is excited again from the trap to the conduction band. In the second trap assisted process the electron is still thermally excited from the valence to a trap, but then it tunnels out as shown by the red arrows. If there is a large trap concentration, this can easily be the dominant current in the PN junction.

In order to model the trap assisted tunneling current we will start the Shockley-Reed-Hall theory and modify the rates to account for tunneling. Figure 5.2 shows the different trap assisted recombination present in a PN junction. Only tunneling to/from the conduction band is

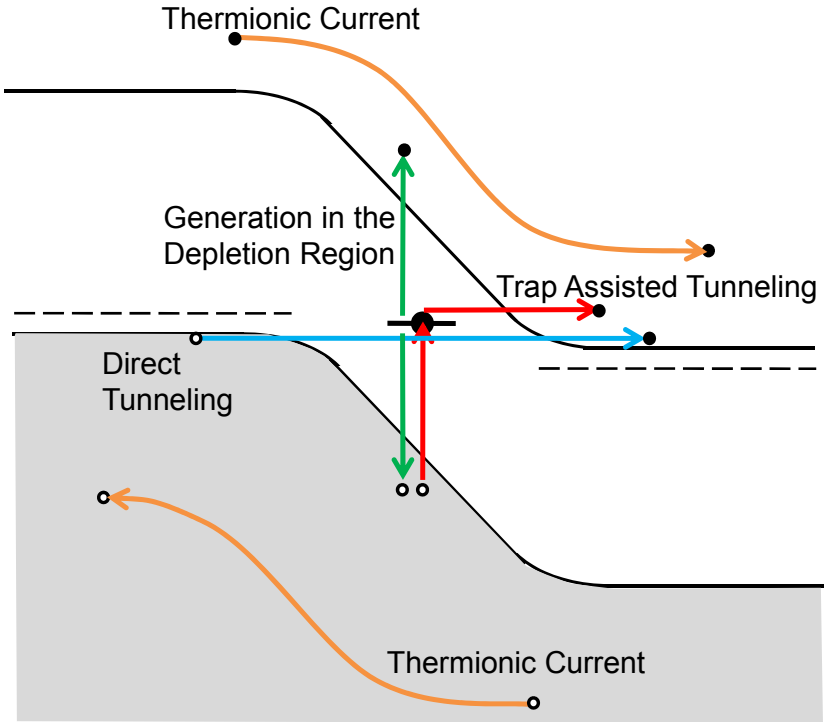


Figure 5.1: The different types of current in a reverse biased PN junction

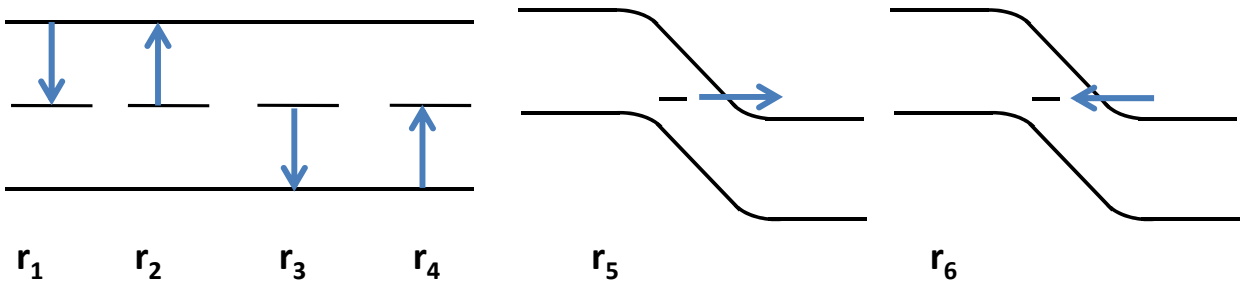


Figure 5.2: Different recombination processes involving a trap are shown. The direction an electron would go in each process is shown.

shown. Tunneling to/from the valence band is also possible. First we start with the SRH rates given below[51]:

$$r_1 = v_{th}\sigma N_t n\left(1 - f(E_t)\right) \tag{5.1.1}$$

$$r_2 = v_{th}\sigma N_t n_i e^{(E_t - E_i)/k_b T} f(E_t) \tag{5.1.2}$$

$$r_3 = v_{th}\sigma N_t p f(E_t) \tag{5.1.3}$$

$$r_4 = v_{th}\sigma N_t n_i e^{(E_i - E_t)/k_b T}\left(1 - f(E_t)\right) \tag{5.1.4}$$

To calculate the rate of tunneling out of a trap, r_5, we start with the rate of thermoionically escaping from the trap, r_2, and appropriately modify it. Since the electron no longer needs to overcome a thermal barrier, the escape rate is increased by the Boltzmann factor, $\exp\left((E_C - E_t)/k_B T\right)$. However, it is also decreased by the tunneling probability, T. The total rate is:

$$\begin{aligned} r_5 &= v_{th}\sigma N_t n_i e^{(E_c - E_i)/k_b T} f(E_t)\times T \\ &= v_{th}\sigma N_t n_i e^{E_G/(2k_b T)} f(E_t)\times T \end{aligned} \tag{5.1.5}$$

To calculate the rate of tunneling into a trap, r_6, we start with r_1, and modify it appropriately.

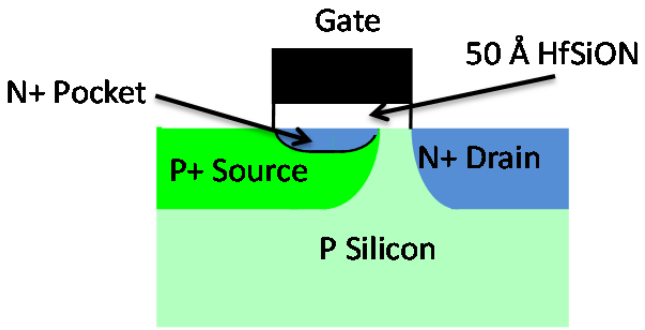


Figure 5.3: We model the trap assisted tunneling in this pocket based TFET. The tunneling occurs between the P+ source and the N+ Pocket

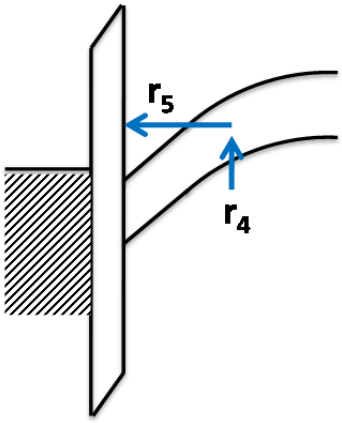


Figure 5.4: The trap assisted tunneling process shown dominates the Id-Vg characteristics of the pocket TFET

In this case the electron density, n_{NL}, is nonlocal and increases to the bulk value where the electron tunneling process starts. However, the rate is also decreased by the tunneling probability, T. The total rate is:

$$r_6 = v_{th}\sigma N_t n_{NL}\left(1 - f(E_t)\right)\times T \tag{5.1.6}$$

Using these rates we can now model a trap assisted tunneling process. A more detailed trap assisted tunneling model is used in these papers [52-54] for diodes, but our simpler model can capture the essential physics in trap limited TFETs. Consider the pocket based TFET shown in Figure 5.3 that was fabricated by Pratik Patel. In this transistor, the tunneling occurs between the N+ pocket and P+ source. Figure 5.4 shows the trap process that dominated the experimental results.

To model the current we need to know the tunneling probability and the energy of the trap level. The tunneling probability can be found by the WKB approximation for a triangular barrier and is given by[13]:

$$T = e^{\frac{-4(2m^*)^{1/2}*(E_c - E_t)^{3/2}}{3\hbar q\vec{E}_S}} \tag{5.1.7}$$

Now we need to know the electric field across the junction. Since we are modeling a vertical tunneling device, we can use the same approximations used in Section 2.2. Namely, we assume the electric field across the oxide is the same electric field in the semiconductor:

$$\vec{E}_S = \frac{\varepsilon_{OX}}{\varepsilon_S}\times\frac{V_G - V_T}{T_{OX}} \tag{5.1.8}$$

Next we assume that there is a uniform distribution of traps in the band gap and find the energy at which the trap assisted tunneling current is maximized. The maximum trap assisted tunneling current will occur when r4 = r5 and the probability of the trap being occupied is 50%. Given a particular voltage we can solve for the trap energy by solving for r4 (E_t)= r5 (E_t). Thus r4=r5 is the rate per unit volume at which carriers are generated through trap assisted tunneling. To get a current we only need to multiply by q and by the volume:

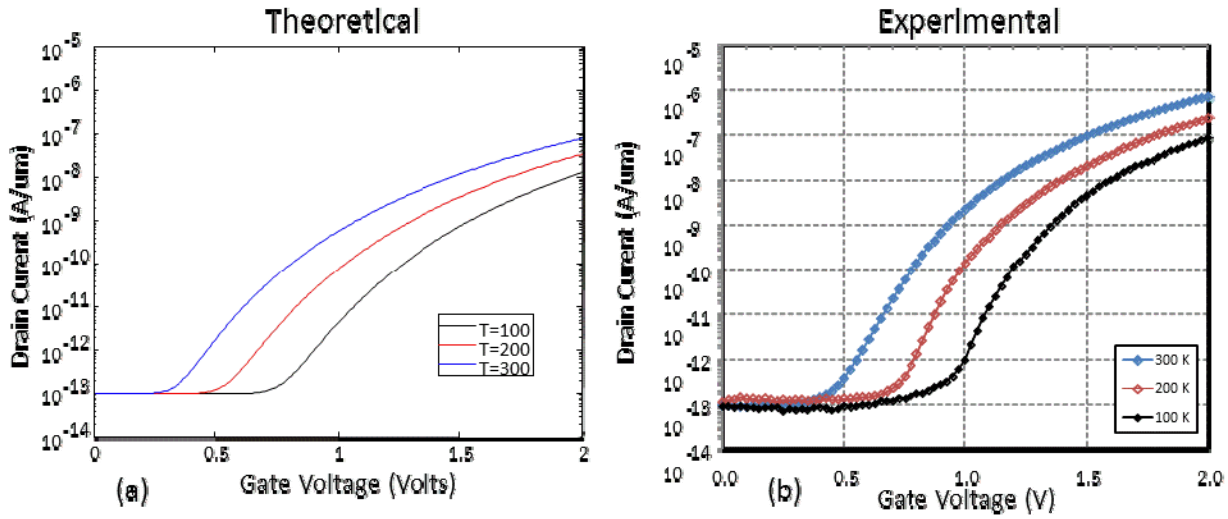


Figure 5.5: (a) The modeled trap assisted tunneling current for the pocket TFET is shown as a function of temperature. (b) The experimental Id-Vg curve as a function of temperature is shown (from Pratik Patel, unpublished).

$$J = q \times V \times r_4(E_t) \tag{5.1.9}$$

Finally we can apply this model to the device shown in Figure 5.3. The key parameters used are $N_t = 10^{12}$ /cm^2, $\sigma = 1$ nm^2, and the background leakage = 10^{-13} A/um. The resulting simulated current as a function of temperature is shown in Figure 5.5(a). The experimental curves (courtesy of Pratik Patel) are shown in Figure 5.5(b). The agreement between our simple model and the experimental curves is not perfect, but several key features are reproduced. The biggest success of the model is that it reproduces the large threshold voltage shift with temperature. It also qualitatively reproduces the shape of the Id-Vg curve. It is possible that the increased experimental on-state conductivity is partly due to the benefits of tunneling to a confined state as discussed in Chapters 6 and 7.

The strong temperature dependence of trap assisted tunneling can also be seen in the reverse bias current of trap limited diodes [54]. Reducing the temperature can reduce the current by orders of magnitude as the thermal part of the trap assisted tunneling process is suppressed. In forward bias, the trap assisted tunneling current (typically the "excess" current in the valley of an Esaki diode) has a weak temperature dependence. This is because carriers fall into the traps and give off energy instead of absorbing energy.

As seen from Figure 5.5 the subthreshold slope for the trap assisted tunneling process is fairly gradual. In fact, it will always be worse than direct tunneling if there is a uniform density of trap states in energy. This is because the trap assisted process will allow current to flow when it should have been cut off by the band edges. However, it may be possible to use traps that are localized in energy to reduce the effective band gap and get a steeper turn on.

5.2 Contact Broadening / Source to Drain Tunneling

Having a contact or a drain near the tunneling junction can cause the energy levels to broaden and reduce the sub-threshold slope. The level broadening will only occur at energies

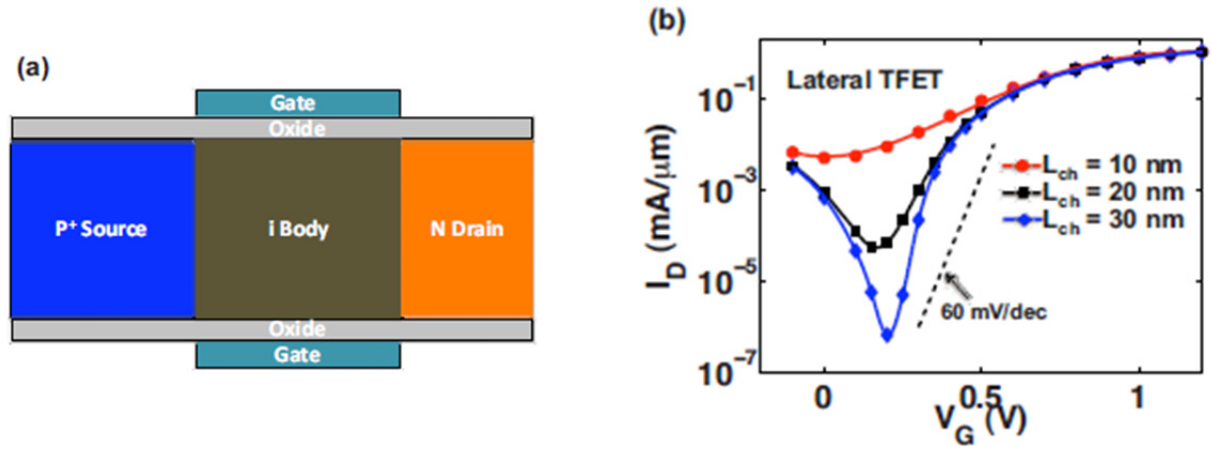


Figure 5.6: (from Ganapathi 2010) As the channel length is reduced the direct source to drain tunneling limits the subthreshold slope.

that are allowed in the contact. (This can be verified by considering an NEGF model. The contact self-energy is non-zero only at energies that are allowed in the contact.) The entire level broadening process can be modeled as direct tunneling to the contact. This means that an electron can directly tunnel to a nearby metallic contact or it can directly tunnel from the source to the drain. The size limitations posed by direct source to drain tunneling will be worse in TFETs than in traditional CMOS. Since TFETs will have a steeper subthreshold slope, the direct source to drain tunneling will need to be suppressed more and so a longer channel will be needed. This will pose a direct limit on the maximum scaling of TFETs. The dependence of the sub-threshold slope on the channel length can be seen in a simulation paper by Ganapathi et all [55] and is reproduced in Figure 5.6. A basic PIN TFET as shown in Figure 5.6(a) was simulated. As the channel length is decreased from 30 nm to 10 nm the subthreshold slope progressively gets worse as seen in Figure 5.6(b).

5.3 Graded Junctions/Poor Electrostatics

Another major concern in designing TFETS is to ensure that the entire tunneling junction turns on at once. If different regions turn on at different biases, the overall I-V curve will be smeared out and the subthreshold slope will not be very steep. These types of effects are analyzed in detail in Pratik Patel's dissertation[56] and so we have just reproduced a couple key figures in Figure 5.7. The simulated structure is shown in Figure 5.7(a). The tunneling occurs between the P^+ source and an N^+ pocket. As indicated by $V_{ov,1}$ and $V_{ov,2}$ there are two possible tunneling paths. The vertical tunneling path corresponding to $V_{ov,1}$ is the desired tunneling path and the lateral tunneling path corresponding to $V_{ov,2}$ is an undesired parasitic tunneling path. If the parasitic tunneling path turns on prior to the desired tunneling path, the subthreshold slope will not be very steep. This is shown in Figure 5.7(b). If the pocket is misaligned, the parasitic tunneling paths will dominate. Figure 5.7(c) shows what happens if the pocket doping profile is

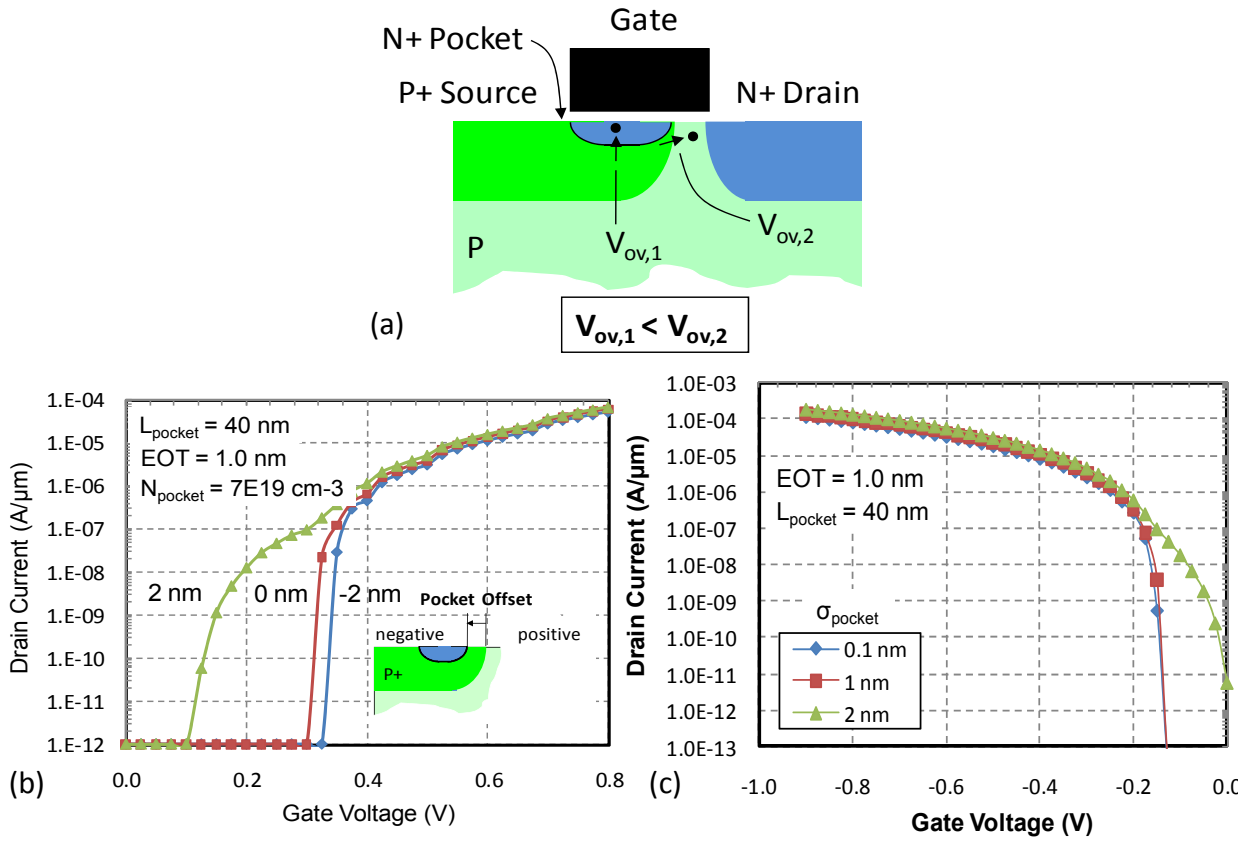


Figure 5.7: (from Patel 2010) (a) A pocket source TFET was simulated and shown to very sensitive to the exact doping profile. (b) The drain current as a function of varying offsets is shown. If designed incorrectly, a gradual lateral tunneling process can dominate the Id-Vg characteristic. (c) If the pocket doping is not abrupt, different regions have different threshold voltages and so the overall turn on is very gradual.

not abrupt. The lateral straggle of the dopants allows for parasitic tunneling paths to dominate and the subthreshold slope loses its steepness.

5.4 Conclusion

As we saw in this chapter there are a variety of effects that can limit the performance of a tunneling field effect transistor. Fortunately, they can be engineered around if one is cognizant of the limitations. Focusing on high quality gate interfaces and material quality will suppress trap assisted tunneling. Using reasonable channel lengths will avoid direct source to drain tunneling (while accepting some more stringent scaling limits). Finally, simulating and correctly designing the electrostatics can avoid any problems with parasitic tunneling paths.

Chapter 6: Pronounced Effect of pn-Junction Dimensionality on Tunnel Switch Sharpness

6.1 INTRODUCTION

Now that we have considered the non-idealities that can affect a tunneling switch, we can look at what impact the ideal density of states will have on a tunneling switch and how we would design an ideal density of states switch.

The density of states turn-on is illustrated in Figure 6.1. If the conduction and valence band do not overlap, no current can flow. Once they do overlap, there is a path for current to flow. This band overlap turn-on has the potential for a very sharp On/Off transition that is much sharper than that which can be achieved by modulating the tunneling barrier height or thickness[7]. If the band edges are ideal, one might expect an infinitely sharp turn on when the band edges overlap. We will find that in a typical 3d-3d bulk pn junction, the nature of the turn on is only quadratic in the control voltage. A sharper density-of-states occurs if the dimensionality on either side of the pn junction is reduced. In specifying a pn junction it is also necessary to specify the dimensionalities of p, and of n regions. We count nine different possible pn junction dimensional combinations, as shown in Figure 6.2.

In the following sections we analyze each of the nine cases, in the following sections: II. 1d-1d; III. 3d-3d; IV. 2d-2d_{edge}; V. 0d-1d; VI. 2d-3d; VII. 1d-2d; VII. 0d-0d; IX. 2d-2d_{face}; X. 1d-1d_{edge}. We ask which are promising for adaptation into a TFET[8], or for a new generation of Backward Diodes?

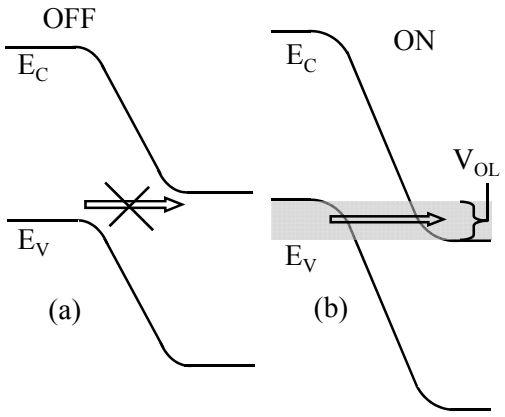


Figure 6.1: (a) No current can flow when the bands do not overlap. (b) Once the bands overlap, current can flow. The band edges need to be very sharp, but density of states arising from dimensionality also plays a role.

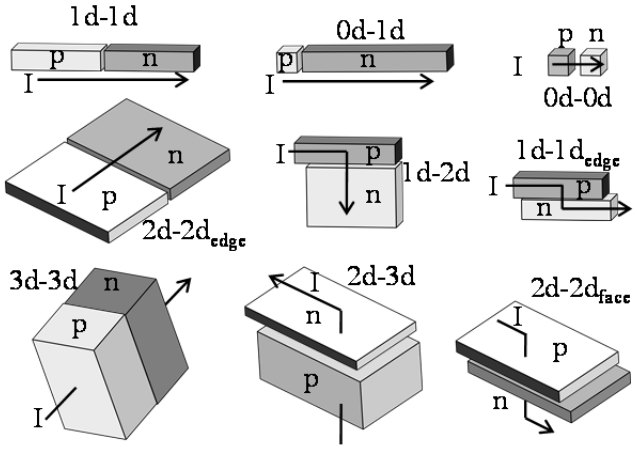


Figure 6.2: We identify here the nine distinct dimensionality possibilities that we believe can exist in pn junctions. Each of the different tunneling pn junction dimensionalities shown have different turn on characteristics

6.2 1D-1D POINT JUNCTION

A 1d-1d point pn junction describes tunneling[9] within a nanowire or carbon nanotube junction as schematically represented in Figure 6.3(a). Tunneling is occurring from the valence band p-side to the conduction band n-side. The gate is not shown as there are many possible gate geometries. The band diagram across this junction is given by Figure 6.3(d).

In analyzing all of the devices, we consider a direct gap semiconductor with a small gate bias. In particular we consider the regime near the band overlap turn-on where a small change in gate voltage (k_bT/q or less) will result in a large change in the density of states but only a small change in the dimensionless tunneling probability. Consequently we assume that the tunneling probability is roughly a constant, T_{device}, and will not change significantly for small changes in the gate voltage.

We also define V_{OL} to be the overlap voltage between the conduction and valence bands as shown in Figure 6.1(b). In a backward diode structure this would be related to the reverse bias. In a transistor structure, this would typically be related to the gate voltage, V_G, and the

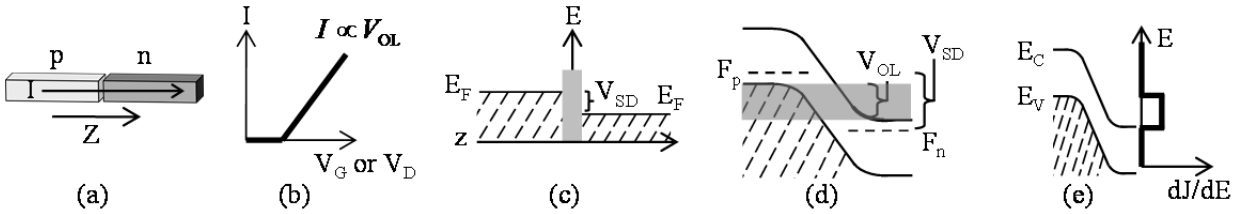


Figure 6.3: Various characteristics of a 1d-1d point overlap junction. (a) The pn junction is oriented in the Z-direction. (b) Linear I-V of the junction (c) Energy versus position for a typical 1d tunnel barrier. (d) Band diagram for the tunnel pn junction showing that the relevant voltage is the overlap voltage and not the source drain voltage. (e) The differential current per unit energy is constant across the tunneling region.

source drain bias voltage, V_{SD}. In order to keep the analysis as simple and general as possible we will use the band overlap voltage, V_{OL} in all of the analyses instead of V_G or V_{SD}.

The 1d nanowire current can be derived as an adaptation of the normal quantum of conductance, $2q^2/h$, approach. The band diagram for the typical quantum of conductance is shown in Figure 6.3(c). The current flow is controlled by the difference in the Fermi levels, which is the source drain voltage, V_{SD}, as shown. Current is given by charge × velocity × density of 1d states. Furthermore, the differential current, dJ/dE, that flows at any given energy is the same at all energies. This arises because the energy dependence of the velocity and 1d density of states exactly cancel, such that current is the same regardless of the energy. This results in a current I, controlled by quantum conductance where $(I=2q^2/h)\times V_{SD}\times \mathcal{T}_{device}$, where $\mathcal{T}_{device}$ is the tunneling probability.

Now to properly consider the transition from conduction band to valence band we look at the band diagram given in Figure 6.3(d). Initially, we consider the situation shown in Figure 6.3(d), where the valence band on the p-side of the junction is completely full and the conduction band on the n-side is completely empty. This would correspond to non-degenerate doping, $V_{SD}>k_bT/q$ and $V_{SD}>V_{OL}$.

As shown in Figure 6.3(d), the band edges cut off the number of states that can contribute to the current. Unlike a single band 1d conductor, the overlap voltage V_{OL} determines the amount of current that can flow. Nevertheless, as shown in Figure 6.3(e), dJ/dE is still independent of energy and is equal to q/h. This arises because the exact same energy dependence cancellation between the velocity and density of states still occurs on both sides of the 1d pn junction.

Thus the 1d pn junction will conduct with a quantum of conductance times the tunneling probability, with the relevant voltage being the overlap voltage.

$$I_{1d-1d} = \frac{2q^2}{h}\times V_{OL}\times \mathcal{T}_{device} \tag{6.2.1}$$

Since a long-range goal is a powering and switching voltage $<k_bT/q$, let us consider the case $V_{SD}<k_bT/q$. To account for the small voltage we need to multiply by the Fermi occupation difference $(f_c - f_v)$. In this small bias regime everything of interest occurs within a k_bT or two of energy. Consequently we can Taylor expand $(f_c - f_v)$:

$$f_c = \frac{1}{e^{(E-E_{Fc})/k_bT}+1} \tag{6.2.2}$$

$$f_c - f_v \approx \frac{(E_{Fc}-E_{Fv})}{4k_bT} \approx \frac{qV_{SD}}{4k_bT} \tag{6.2.3}$$

Thus the ultimate effect of the small differential Fermi occupation factors is to multiply the low temperature current by the factor $qV_{SD}/4k_bT$. We can therefore write a conductance for small source drain biases:

$$G_{1d-1d} = \frac{2q^2}{h}\times V_{OL}\times \mathcal{T}_{device}\times \frac{q}{4k_bT} \tag{6.2.4}$$

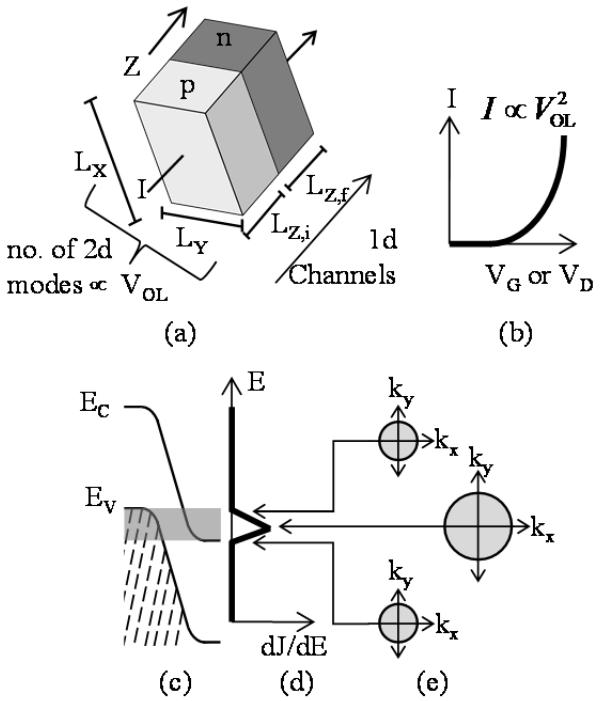


Figure 6.4: Various characteristics of a 3d-3d bulk junction. (a) Schematic representation of the pn junction (b) Quadratic I-V threshold behavior of the junction (c) Band diagram (d) The differential current per unit energy is proportional to the number of transverse states (e) The transverse k states that contribute to the current at various energies

This is true for all of the following devices to be considered in this article, as well. Thus we will continue to make the approximation that the valence band is full and the conduction band is empty when calculating the potential current flow, and then add the effect of the partial Fermi occupation functions afterwards. A more rigorous derivation of the tunneling current and the effect of the Fermi functions are given in Appendix A and B (Section 6.16 and 6.17).

6.3 3D-3D BULK JUNCTION

A 3d-3d junction simply means a pn junction or heterojunction where there is a bulk semiconductor on either side of the sample. A generalized schematic of the tunneling junction only is shown in Figure 6.4(a). The band diagram across this junction is given by Figure 6.4(c).

The 3-d bulk current can be derived from a few simple considerations. The junction is a large 2d surface and can be considered to be a 2d array of 1d channels. The 2d array is defined by the transverse k-states that can tunnel. Each 1d channel is equivalent to the 1d-1d case described in the previous section and will conduct with a quantum of conductance times the tunneling probability. The differential current density can therefore be written as:

$$\partial \mathrm{I} = \mathrm{N}_{\perp \mathrm{states}} \times \frac{2\mathrm{q}}{h} \times \mathcal{T}_{\mathrm{device}} \times \partial \mathrm{E} \tag{6.3.1}$$

The number of transverse states is the number of k-states within the maximum transverse energy at a given energy. The transverse energy is limited by the closest band edge and peaks in the middle of the overlap. This is shown in Figure 6.4(e). The differential current density is given by Figure 6.4(d). Integrating over the overlap gives:

$$I_{3d-3d} = \frac{1}{2}\left(\frac{Am^*}{2\pi\hbar^2}\times\frac{qV_{OL}}{2}\right)\times\frac{2q^2}{h}V_{OL}\times\mathcal{T}_{device} \tag{6.3.2}$$
$$= \text{No. of 2d Channels}\times\text{1d Conductance}$$

for large $V_{OL}>k_bT/q$ where A is the area of the junction.

For small $V_{OL}<k_bT/q$ the conductance can be written as:

$$G_{3d-3d} = \frac{1}{2}\left(\frac{Am^*}{2\pi\hbar^2}\times\frac{qV_{OL}}{2}\right)\times\frac{2q^2}{h}V_{OL}\times\mathcal{T}_{device}\times\frac{q}{4k_bT} \tag{6.3.3}$$

Thus for very small biases the current is quadratic in the overlap voltage as shown in Figure 6.4(b). This is the exact same result that comes from taking the appropriate limits of Kane's tunneling theory[22].

In the Appendix, we also formally derive this result in a different manner using the transfer Hamiltonian method[57-60]. We do this as an alternative to employing the more modern channel conductance approach. The transfer Hamiltonian method was first used by Oppenheimer to study the field emission of hydrogen[60]. It was then expanded by Bardeen[57] for tunneling in superconductors and then the case of independent electrons was considered by Harrison[59]. The transfer Hamiltonian method in the Appendix is just an application of Fermi's golden rule with a clever choice of states and perturbing Hamiltonian.

6.4 2D-2D EDGE JUNCTION

A 2d edge overlapped junction describes a junction where the tunneling occurs along a line separating p and n regions within in a 2d confined surface. The junction is schematically represented in Figure 6.5(a). This could be represented by a case where tunneling occurs within a thin inversion region near a surface or within an ultra-thin body device.

The derivation of the current is almost identical to the 3d-3d case, except that instead of having a 2d array of 1d channels we now have a 1d array of 1d channels. Therefore the current is:

$$I_{2d-2d,edge} = \frac{2}{3}\left(\frac{L_X\sqrt{m^*}}{\pi\hbar}\times\sqrt{qV_{OL}}\right)\times\left(\frac{2q^2}{h}\times V_{OL}\times\mathcal{T}_{device}\right) \tag{6.4.1}$$

$$=\text{No. of 1d Channels}\times\text{1d Conductance}$$

for large $V_O>k_bT/q$ where L_X is the length of the junction.

For small $V_{OL}<k_bT/q$ the conductance can be written as:

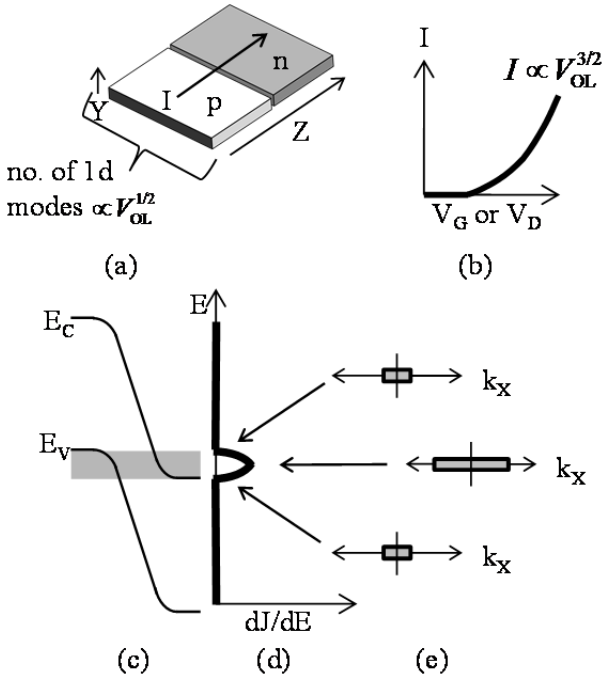


Figure 6.5: Various characteristics of a 2d-2d edge overlap junction. (a) Schematic representation of the junction (b) Power law I-V characteristic of the junction (c) Band diagram (d) The differential current per unit energy dJ/dE is proportional to the number of transverse states (e) The transverse k-states that contribute to the current at various energies

$$G_{2d-2d,edge} = \frac{2}{3}\left(\frac{L_X\sqrt{m^*}}{\pi\hbar}\times\sqrt{qV_{OL}}\right)\times\left(\frac{2q^2}{h}\times V_{OL}\times\mathcal{T}_{device}\right)\times\left(\frac{q}{4k_bT}\right) \qquad (6.4.2)$$

Thus for very small gate biases the current is proportional to $V_{OL}^{3/2}$ as shown in Figure 6.5(b). Similar to the 3d case, the number of transverse states that can tunnel varies with energy and needs to be properly integrated. As shown in Figure 6.5(d) and (e), the differential current density is proportional to the number of states that can tunnel. The number of 1d states is proportional to the square root of the energy and so the differential current density follows a square root with respect to energy.

6.5 0D-1D JUNCTION

A 0d to 1d junction represents tunneling from a quantum dot to a nanowire as shown in Figure 6.6(a). Our main goal in analyzing this case is to provide the basis for analyzing higher dimensionality systems such as a 2d-3d or 1d-2d junctions. Consequently we consider two different 0d-1d systems. First we will assume that there is an electron in the quantum dot and find the rate at which it escapes into the end of a 1d wire. In reality, there is no way to electrically contact the quantum dot. Therefore we consider a more realistic situation that includes the need to couple current into the dot. This case essentially evolves into a single electron transistor (SET) as shown in Figure 6.6(e) and (g).

The rate at which an electron escapes from the quantum dot into a nanowire is given by the field ionization of a single state such as an atom. In Gamow's model of alpha particle decay[61], the particle is oscillating back and forth in its well and it attempts to tunnel on each round trip oscillation. If the dot has a length of L_Z along the tunneling direction, the electron will travel a distance of $2L_Z$ between tunneling attempts. Its momentum is given by $p_Z=mv_Z=\hbar k_Z$ where $k_Z=\pi/L_Z$ in the ground state. Using $E_Z=\hbar^2 k_Z^2/2m$, the time between tunneling attempts is $\tau=2L_Z/v_z=h/2E_z$. The tunneling rate per second is $R=(1/\tau)\times\mathcal{T}_{device}$. This can be converted to a current by multiplying by the electron charge, and a factor 2 for spin to give:

$$I=\frac{4q}{h}\times E_Z\times\mathcal{T}_{device} \tag{6.5.1}$$

This is the same result that one obtains from the transfer Hamiltonian method outlined in the Appendix. It also assumes a large source drain bias, as usual. These simple considerations predict a constant current as soon as the bands overlap as shown in Figure 6.6(b). As seen from the band diagram in Figure 6.6(c), all tunneling occurs at the single confined energy.

To include coupling into the dot, we add a second nanowire to supply current, as shown in Figure 6.6(e) or (g). We assume that the second nanowire has the same tunneling probability/coupling strength to the quantum dot as the original one. The new band diagram is shown in Figure 6.6(f) or (h), and resembles that of a "single electron transistor"[62]. As in Figure 6.6(c), the tunneling occurs at a single energy and will result in a constant current once the bands overlap. The tunneling event out of the dot follows sequentially after tunneling in. Therefore the total current tunneling transport rate is halved $(1/2\tau)\times\mathcal{T}_{device}$, and the current is cut in half:

$$I=\frac{2q}{h}\times E_Z\times\mathcal{T}_{device} \tag{6.5.2}$$

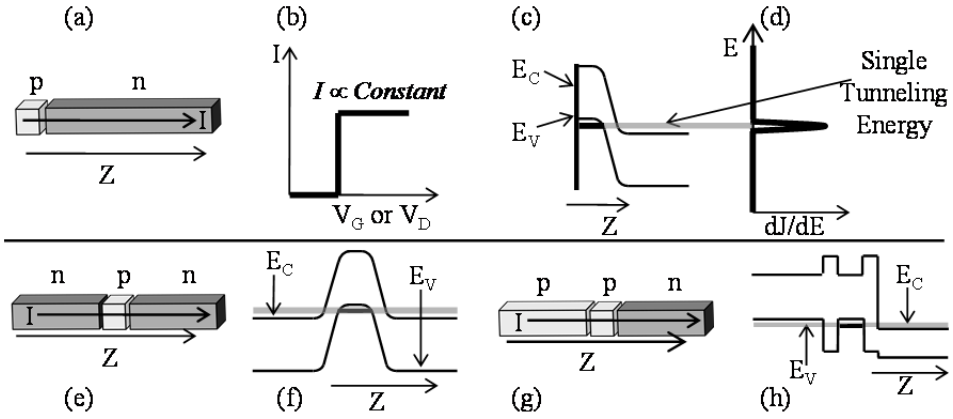


Figure 6.6: Various characteristics of a 0d-1d end junction. (a) Schematic representation of the junction. (b) Step function I-V of the junction. (c) Band diagram. (d) Differential current per unit energy. (e) More realistic 1d single electron transistor (SET) structure. (f) Band diagram corresponding to the more realistic SET. (g) Alternate SET structure with a p-type contact (h) Band diagram corresponding to the alternate SET

This applies to a filled source wire, and an empty drain wire. If there is partial Fermi-Dirac occupation on both sides we can define a conductance based on the Fermi Level difference:

$$G = \frac{2q}{h} \times E_Z \times \mathcal{T}_{device} \times \frac{q}{4k_b T} \tag{6.5.3}$$

While our simple model assumes a perfectly sharp level and therefore a step turn-on in the current, in reality the level will be broadened and the turn-on threshold will assume the shape of the level broadening. This broadening can be extrinsically caused by any inhomogeneities in the lattice such as defects, dopants, or phonons. Even without these effects, simply coupling to the dot to the nanowires will already cause a significant amount of broadening[63]. In the simplest model, the density of states in the level broadens to form a Lorentzian with a full width in energy at half maximum of γ. We seek to find γ, since that will determine the inherent broadening at threshold. We start with the escape rate of electrons from the dot $=(4E_z/h)\times\mathcal{T}_{device}$, where the escape rate is doubled, since there are two extra wires that can capture the escaped electron. Multiplying by $\hbar$ leads to a broadening:

$$\gamma = (2/\pi) \times E_Z \times \mathcal{T}_{device} \tag{6.5.4}$$

which allows us to write the conductance as $G = (q^2/h) \times \gamma \times (\pi/4k_bT)$.

We can now define a Figure-of-Merit for 0d-1d switches:

$$\frac{\text{Conductance}}{\text{Broadening}} = \frac{G}{\gamma} = \frac{\pi q^2/h}{4k_b T} \tag{6.5.5}$$

which can be rewritten as:

$$\frac{G}{q^2/h} = \frac{\pi \gamma}{4k_b T} \tag{6.5.6}$$

This means that a steep switch with broadening $\gamma << k_bT$, must come as the expense of a conductance much less than the conductance quantum $G << (q^2/h)$.

6.6 2D-3D JUNCTION

A 2d-3d tunneling junction is typical in vertical[3] TFET's where the tunneling occurs from the bulk to a thin confined layer under the gate. The thin layer can either be a thin inversion layer or a physically separate material. A generalized schematic of this tunnel junction is shown in Figure 6.7(a). Here the z-axis is rotated 90° from what one would usually expect so that the axes remain parallel between figures.

The derivation for this case is very similar to the 3d-3d case. As in that section, the junction is a large 2d surface and can be considered to be a 2d array of 1d tunneling problems. However, this case does not represent the typical 1d quantum of conductance. In this case, the 1d problem is better described by the field ionization of a single state such as an atom as described in the 0d-1d section. We simply multiply that result by the number of 2d channels to get a current of:

$$I = \text{No. of 2d channels} \times \text{1d field ionization}$$

$$I = \left(\frac{\text{Am}}{2\pi\hbar^2} \times \frac{qV_{OL}}{2} \right) \times \left(\frac{4q}{h} \times E_Z \times \mathcal{T}_{\text{device}} \right) \tag{6.6.1}$$

for large $V_{OL}>k_bT/q$.

For small $V_{OL}<k_bT/q$ the conductance can be written as:

$$G_{\text{2d-3d}} = \left(\frac{\text{Am}}{2\pi\hbar^2} \times \frac{qV_{OL}}{2} \right) \times \left(\frac{4q}{h} \times E_Z \times \mathcal{T}_{\text{device}} \right) \times \left(\frac{q}{4k_bT} \right) \tag{6.6.2}$$

Here, E_Z is the confinement energy of the 2d layer. This is the exact same result that comes from the transfer Hamiltonian method in the Appendix so long as we also assume that the confined electron is in the ground state such that $k_Z=\pi/L_Z$. Thus for very small V_{SD}, or small gate biases, the current is linear in V_{OL} as shown in Figure 6.7(b). Compared to the bulk 3d-3d case, confining one side of the junction resulted in the replacement of qV_{OL} with $4E_Z$.

To justify the number of transverse states that we have included we need to look more closely at the tunneling process. The band diagram is shown in Figure 6.7(c). Since the tunnel rate from the 2d quantum well states is constant, the differential current per unit energy is exactly proportional to the number of states that tunnel at any given energy. The number of states that tunnel is equal to the number of transverse states, and since the 2d density of states is independent of energy, the differential current per unit energy is also a constant. Figure 6.7(e)

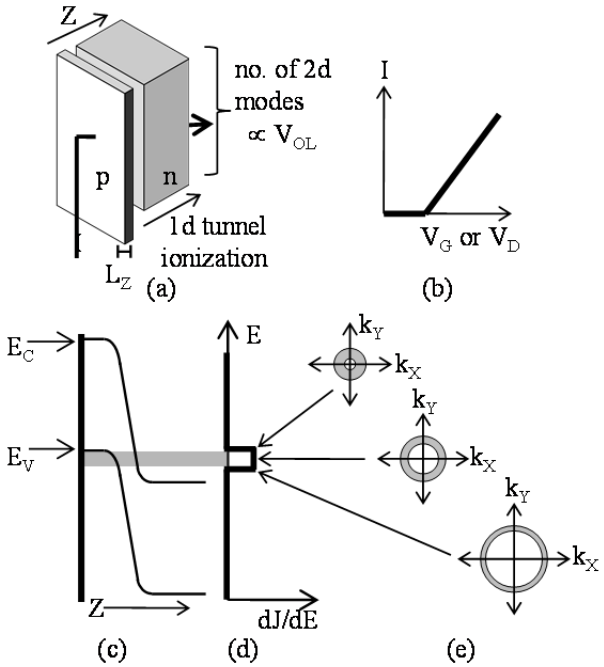


Figure 6.7: Various characteristics of a 2d-3d junction. (a) Schematic representation of the junction (Z-axis rotated 90° from the usual.) (b) Linear I-V threshold response of the junction (c) Band diagram (d) The differential current per unit energy is proportional to the number of transverse states (e) The transverse k-states that contribute to the current at various energies

shows that the states that tunnel at any given energy map out a ring in k-space of constant area. Oddly, the figures also show current flowing in only the upper half of the energy overlap region, at low kinetic energy in the valence band. At higher kinetic energy in the valence band, the transverse momentum becomes too large to fit into the small region of k-space on the conduction band side. Thus only valence band states with a kinetic energy up to $V_{OL}/2$ will tunnel.

Current can flow in along the transverse direction as shown in Figure 6.7(a). Other methods can also be considered for making electrical contact.

6.7 1D-2D JUNCTION

A 1d-2d junction describes tunneling between the edge of a nanowire and a 2d sheet as shown in Figure 6.8(a). The derivation for this case is very similar to the 2d-3d case. The only difference is that instead of a 2d array of 1d tunneling, we now have a 1d array of 1d tunneling. Thus the current is:

$$\begin{aligned} I_{1d\text{-}2d} &= \text{no. of 1d channels} \times \text{1d tunnel ionization} \\ I_{1d-2d} &= \left(\frac{L_X}{\pi\hbar} \times \sqrt{qm^* V_{OL}}\right) \times \left(\frac{4q}{h} \times E_Z \times \mathcal{T}_{device}\right) \end{aligned} \tag{6.7.1}$$

for large $V_{OL}>k_bT/q$.

For small $V_{OL}<k_bT/q$ the conductance can be written as:

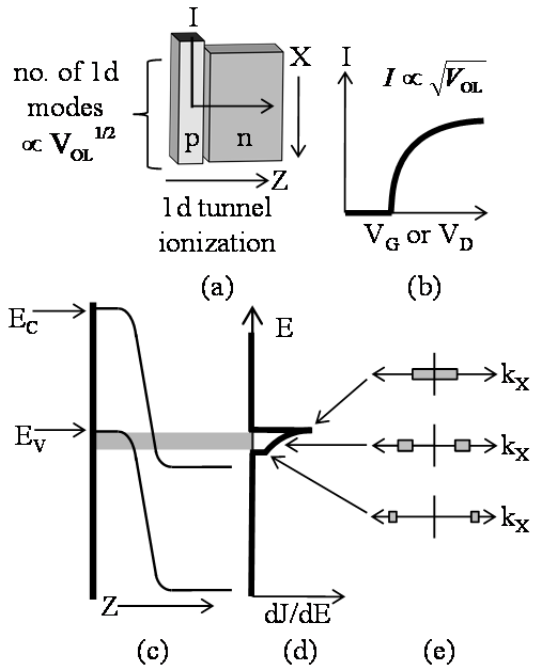


Figure 6.8: Various characteristics of a 1d-2d junction. (a) Schematic representation of the junction (b) Square root I-V threshold of the junction (c) Band diagram (d) The differential current per unit energy is proportional to the number of transverse states (e) The transverse k-states that contribute to the current at various energies

$$G_{1d-2d} = \left(\frac{L_X}{\pi\hbar}\times\sqrt{qm^*V_{OL}}\right)\times\left(\frac{4q}{h}\times E_Z\times\mathcal{T}_{device}\right)\times\left(\frac{q}{4k_bT}\right) \tag{6.7.2}$$

where, E_Z is the confinement energy along z-axis of the 1d layer. This is exactly the same result that comes from the transfer Hamiltonian method in the Appendix as long as we also assume that the confined electron is in the ground state such that $k_Z=\pi/L_Z$ as before. Thus for very small V_{OL} the current is proportional to sqrt(V_{OL}) as shown in Figure 6.8(b). In addition, comparing to the 2d-2d edge overlap formula, confining one side of the junction resulted in the replacement of qV_{OL} with $3E_Z$.

As in the 2d-3d case, current flows in only the upper half of the energy overlap region at low kinetic energy in the valence band, due to conservation of transverse momentum. This is indicated by the shaded part of the band diagram in Figure 6.8(c). At higher kinetic energy in the valence band, the transverse momentum becomes too large to fit into the small region of k-space on the conduction band side. Therefore, we included transverse states only up to half the overlap voltage. In this case, the 1d density of states varies with energy and so the current that tunnels at different energies is different. This is shown in Figure 6.8(d). Figure 6.8(e) shows which transverse states contribute at each energy.

6.8 0D-0D QUANTUM DOT TUNNELING

This case simply represents tunneling from a filled valence band quantum dot to an empty conduction band quantum dot. It is schematically represented in Figure 6.9(a). In order to create a meaningful device the quantum dots need to be coupled to contacts, to pass current in and out of the device. This coupling is indicated by the tunnel junctions in Figure 6.9(a). By itself this case is not very interesting, but it will be useful for describing some of the other cases. As with the 0d-1d case, we will consider two different 0d-0d systems. Initially we will ignore the effects of the contacts to the dots and then we will include the effects of contacting the quantum dots in order to make a realistic device.

If we have two isolated quantum dots that are coupled to each other with an electron in one of the dots, the electron will quantum mechanically oscillate back and forth between the dots. The band diagram for this situation is shown in Figure 6.9(c). The thick lines represent the confined energy levels. In order to calculate the rate at which the electron travels between the two states we need to use time dependent perturbation theory (TDPT). The standard result from TDPT for the transition probability of a two level system subject to a constant perturbation is[64]:

$$P_{i\to f}(t) = 4\frac{\left|M_{fi}\right|^2}{\hbar^2}\frac{\sin^2\left(\omega_{fi}t/2\right)}{\omega_{fi}{}^2} \tag{6.8.1}$$

Where $\omega_{fi} = (E_f - E_i)/\hbar$ and M_{fi} is the transition matrix element between the states as defined in Appendix A (Section 6.16). In the Appendix we derived the current using Fermi's golden rule. Evaluating the matrix element is similar and is given by Eqn. (6.16.12):

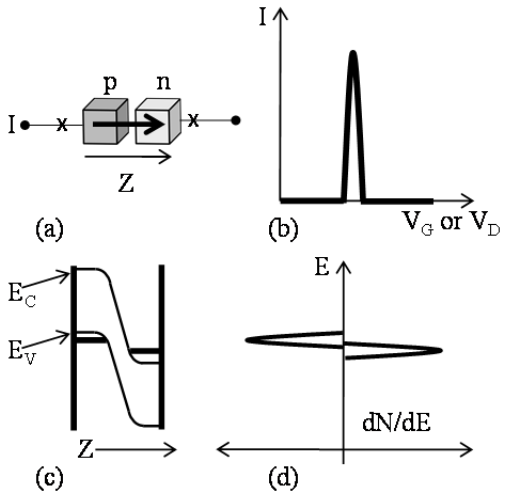


Figure 6.9: The properties of a weak 0d-0d junction that is coupled to electrical contacts. (a) Schematic representation of the junction (b) I-V of the junction (c) Band diagram (d) The density of states per unit energy on left and right sides of the pn junction.

$$\left|M_{fi}\right|^2 = \frac{1}{\pi^2} E_{Z,i} \times E_{Z,f} \times \mathcal{T}_{device} \tag{6.8.2}$$

Plugging Eqn. (6.8.2) into Eqn. (6.8.1) gives:

$$P_{i\to f}(t) = \frac{4}{\hbar^2 \pi^2} E_{Z,i} \times E_{Z,f} \times \mathcal{T}_{device} \times \frac{\sin^2\left(\omega_{fi} t/2\right)}{\omega_{fi}^2} \tag{6.8.3}$$

Here, $E_{Z,i}$ is the confinement energy along the z-axis of the initial dot and $E_{Z,f}$ is the confinement energy along the z-axis of the final dot. Thus the probability that the electron is in the final dot oscillates back and forth and its magnitude is set by the confinement energies in each dot, as given in Eqn. (6.8.3).

Now we consider what happens when the electrical contacts couple to the dot. The levels in the quantum dots will broaden due to the contact coupling and will have a broadened density of states as shown in Figure 6.9(d). When the quantum dots are aligned, current will flow. This results in a single peak in the I-V curve as shown in Figure 6.9(b). The width of the current peak is determined by the contact broadening. For a single level coupled to a single contact the energy broadening will be $\gamma = \hbar/\tau = \hbar\, I/q$. As in Section V, the electron's escape rate is analogous to the attempt frequency of Gamow's theory[61] of alpha decay tunneling and is given by $2E_{Z,i}/h$. The maximum current to the contact per spin state is half of Eqn. (6.5.1): $I=(2q/h)\times E_{Z,i}\times \mathcal{T}_{contact}$., where the contact tunneling transmission is labeled $\mathcal{T}_{contact}$ to distinguish it from the interdot transmission $\mathcal{T}_{device}$. Consequently, the broadening will be given by:

$$\gamma=(1/\pi)\times E_Z \times \mathcal{T}_{contact} \tag{6.8.4}$$

For simplicity, we assume that the confinement energies are the same in both dots: $E_{Z,i} = E_{Z,f} = E_Z$.

The transition probability, Eqn. (6.8.3), resembles a delta function, which when integrated leads to Fermi's Golden Rule: $\text{Rate} = \frac{2\pi}{\hbar} |\, M_{fi} \,|^2 \frac{dN}{dE}$, where dN/dE is the density of final states. In the 0d case, there is only one state, and so the density of states is the inverse of the broadening, γ: $dN/dE = 1/\gamma = \pi/(E_Z \times \mathcal{T}_{contact})$. The transition rate leads to a current:

$$I = 2q \frac{2\pi}{\hbar} |\, M_{fi} \,|^2 \frac{\pi}{E_Z \times \mathcal{T}_{contact}} \tag{6.8.5}$$

$$I = 2q \frac{2\pi}{\hbar} \left[\frac{1}{\pi^2} E_Z \times E_Z \times \mathcal{T}_{device} \right] \frac{\pi}{E_Z \times \mathcal{T}_{contact}} \tag{6.8.6}$$

$$I = \frac{4q}{\hbar} E_Z \times \frac{\mathcal{T}_{device}}{\mathcal{T}_{contact}} \tag{6.8.7}$$

We will show in Section XI that the condition for perturbation theory to be valid requires that the interdot tunneling transmission $\mathcal{T}_{device}$ be less than the square contact tunneling transmission: $\mathcal{T}_{device} << (\mathcal{T}_{contact})^2$. This assures that the interdot current is always less than the maximum contact current. Eqn. (6.8.7) can now be adapted to allow for Fermi occupation of the dots, which leads to a conductance:

$$\sigma = \frac{4q}{\hbar} E_Z \times \frac{\mathcal{T}_{device}}{\mathcal{T}_{contact}} \times \frac{q}{4k_b T} \tag{6.8.8}$$

We can now define a Figure-of-Merit for 0d-0d switches as before:

$$\frac{\text{Conductance}}{\text{Broadening}} = \frac{G}{\gamma} = \frac{8\pi^2 q^2/h}{4k_b T} \times \frac{\mathcal{T}_{device}}{\mathcal{T}_{contact}^2} \tag{6.8.9}$$

which can be rewritten as:

$$\frac{G}{q^2/h} = \frac{2\pi^2 \gamma}{k_b T} \times \frac{\mathcal{T}_{device}}{\mathcal{T}_{contact}^2} \tag{6.8.10}$$

Since the perturbation condition, Eqn. (6.11.4), is that $[\mathcal{T}_{device}/(\mathcal{T}_{contact})^2] < 1$, this places an upper limit on the conductance $G/(q^2/h) < 2\pi^2\gamma/(k_b T)$. This means that a steep switch with broadening $\gamma << k_b T$, must come as the expense of a conductance much less than the conductance quantum $G << (q^2/h)$.

6.9 2D-2D FACE OVERLAP

A 2d-2d area overlapped junction describes tunneling from one quantum well to another through the face of the quantum well. This is different from resonant interband tunnel diodes[65], since the tunneling proceeds from the valence to conduction bands. The junction is schematically represented in Figure 6.10(a). Here the schematic is rotated 90° from what one would usually expect so that the z-axis lines up between figures. This is one of the most

interesting cases as it is the closest to a step function turn-on as illustrated in Figure 6.10(b). The band diagram is shown in Figure 6.10(c).

The step function turn on can be seen by considering the conservation of transverse momentum and total energy. This depicted in Figure 6.10(e). The lower paraboloid represents all of the available states in k-space on the left side of the junction and the upper paraboloid represents the available k-space states on the right side of the junction. In order for current to flow the initial and final energy, and wave-vector k, must be the same and so the paraboloids must overlap. However, as seen in the right part of the figure, they can only overlap at a single energy. Furthermore, the joint density of state pairs between valence and conduction band is a constant in energy. Thus the number of state pairs that tunnel is a constant regardless of the overlap energy as seen in Figure 6.10(d).

The tunneling rate of the valence/conduction band state pairs that transition is different from the 3d-3d bulk case where we had a 2d array of 1d channels. In this case we have a fixed number of 2d states and a completely different 1d problem. The 1d problem now represents tunneling from a single fixed level to another single fixed level as in the 0d-0d case. As seen from the transition probability in the 0d-0d case, Eqn. (6.8.3), the transition probability resembles a delta function as in Fermi's Golden Rule. This will be integrated below, and is independent of the overlap voltage. Thus the total tunneling current is independent of the overlap voltage and will be a step function with respect to the gate voltage.

In order to calculate the amount of current that flows we need to find the transition rate for each state that tunnels and sum over all of the states that tunnel. Every initial state is coupled to only one final state. This is due to conservation of momentum. This means that we can use the 0d-0d result, Eqn. (6.8.3), to describe the transition probability between an initial and final state. Current can flow into each quantum state, along the quantum well, or through the face of the quantum well. We also do not need to externally impose conservation of energy as the 0d-0d result in time dependent perturbation theory is sharply peaked about $E_i=E_f$. Therefore we simply

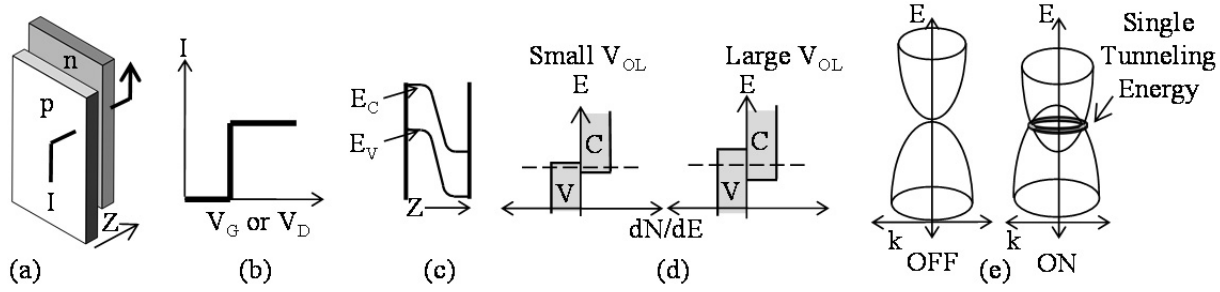


Figure 6.10: Various characteristics of a 2d-2d face overlap junction. (a) Schematic representation of the junction (b) The I-V characteristic is a step function. (c) The band diagram along the tunneling direction shows that the electron is tunneling from one confined sheet to another. (d) Even though the overlap of the density of states increases with increasing overlap voltage, there is only a single energy, indicated by the dotted line, at which the electrons tunnel. (e)There is only a single tunneling energy because of the simultaneous conservation of energy and momentum. The energy versus wave vector paraboloids on each side of the junction only intersect at a single energy.

need to sum Eqn. (6.8.3) over all initial states or final states:

$$\begin{aligned}\sum P_{i\to f} &= \frac{Am}{2\pi\hbar^2}\int_0^\infty P_{i\to f}\,dE \\ &= \frac{Am}{2\pi\hbar^2}\times\frac{2}{h}E_{Z,i}E_{Z,f}\mathcal{T}_{device}\times t\end{aligned} \tag{6.9.1}$$

When evaluating the integral we used equal valence and conduction band mass. Here we see that oscillation in transition probability of individual states is averaged out in summation and that the probability of being in the final state is proportional to time t, as is usual in time-dependent perturbation theory. We also assumed that the bands were sufficiently overlapped and that the lower limit of the integral can be taken to be -∞. This form of Fermi's golden rule is done in Appendix A (Section 6.16).

We can then convert this to a transition rate by taking the time derivative. The transition rate can then be converted to a current density by multiplying by the electron charge and 2 for spin to give a current of:

$$I_{2d-2d,face} = \frac{qmA}{\pi^2\hbar^3}\times E_{Z,i}\times E_{Z,f}\times\mathcal{T}_{device} \tag{6.9.2}$$

for large $V_{SD}>k_BT/q$. For small V_{SD} the conductance depends on the difference in Fermi occupation fractions, and can be written as:

$$G_{2d-2d,face} = \frac{qmA}{\pi^2\hbar^3}\times E_{Z,i}\times E_{Z,f}\times\mathcal{T}_{device}\times\frac{q}{4k_bT} \tag{6.9.3}$$

The main change in going from 3d-3d to 3d-2d is that the energy factor qV_{OL} became E_Z. Likewise, in going from the 3d-2d to 2d-2d the other energy factor qV_{OL} also became E_Z. Thus for each confined side of the junction the relevant energy changes from the overlap energy to the confinement energy. Consequently the 2d case has roughly the same current as a 3d case if the confinement energy E_Z is the same as the overlap voltage qV_{OL}. In practice E_Z is likely to be much larger than qV_{OL}, providing the 2d-2d case with a significant current boost.

Following the joint density of states, the current takes the form of a step function with respect to the gate voltage. This is similar to the step function case of quantum well optical transitions. As soon as the bands overlap, the current immediately turns on. However, contact broadening mechanisms will smear out the step-like turn-on function and this will be discussed later.

6.10 1D-1D EDGE OVERLAP

A 1d-1d edge overlap junction represents two nanowires overlapping each other as shown in Figure 6.11(a). This junction is similar to the 2d-2d area overlap. The current can be found by summing the 0d-0d result over a 1d density of states. Alternatively, the method in the appendices will also give the same result. The resulting current is:

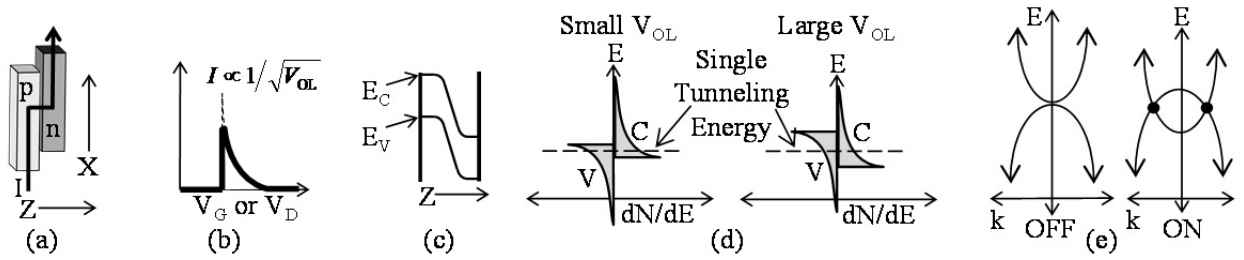


Figure 6.11: Various characteristics of the 1d-1d edge overlap junction. (a) Schematic representation of the junction. (b) I-V curve of the junction. (c) Band diagram along the tunneling direction (d) The 1d density of states at the tunneling energy is different at different overlap voltages (e) There is only a single tunneling energy because of the simultaneous conservation of energy and momentum. The energy versus wave-vector parabolas on each side of the junction only intersect at a single energy.

$$I_{1d-1d,\,edge} = 2\frac{qL_X}{\pi^2\hbar^2}E_{Z,i}\times E_{Z,f}\times\sqrt{\frac{m}{qV_{OL}}}\times\mathcal{T}_{device} \tag{6.10.1}$$

for large V_{OL}.>k_bT/q.

For small V_{OL}<k_bT/q the conductance can be written as:

$$G_{1d-1d,\,edge} = 2\frac{qL_X}{\pi^2\hbar^2}E_{Z,i}\times E_{Z,f}\times\sqrt{\frac{m}{qV_{OL}}}\times\mathcal{T}_{device}\times\frac{q}{4k_bT} \tag{6.10.2}$$

As in the 2d-2d_{face} case the tunneling only occurs at a single energy due to the conservation of momentum and energy as shown in Figure 6.11(e). Since we are now dealing with 1d nanowires, the number of transverse states follows a 1d density of states which follows a $1/\sqrt{V_{OL}}$ dependence. This is illustrated in Figure 6.11(d). This predicts a step function turn on followed by a reciprocal square root decrease. This seemingly implies that the initial conductance will be infinite. However, the contact series resistance will limit the conductance and the broadening associated with the contacts will also limit the peak conductance.

6.11 PERTURBATION TUNNEL TRANSMISSION LIMIT

When a level on the p-side of a junction interacts with a level on the n-side of the junction it is possible for the two levels to interact strongly and repel each other. In most cases this is not a problem, as the interaction between any two *particular* levels goes to zero as the devices get larger and any small amount of level broadening will wash out the level repulsion. In the case of very large contact regions leading to the tunnel junction, the large normalization volume of the wave functions guarantees that individual level repulsion matrix elements are negligible.

In contrast, the 0d-0d, 1d-1d edge overlap and 2d-2d area overlap cases, have finite extent along the tunneling direction, restricting the normalization volume. This means that the tunnel interaction matrix element, $|M_{fi}|$, can take on a large finite value. If this interaction is too

large, the two interacting levels will be strongly coupled and all the perturbation results in this paper will fail. Contrarily, if the level broadening is greater than the level repulsion matrix element, $\gamma > |M_{fi}|$, the level repulsion will be washed out justifying our perturbation approach. The broadening γ is typically caused by coupling to the contacts. It is also possible for various scattering mechanisms to broaden the level.

We will show that if the current is limited by the weak tunneling junction rather than the contact resistances, we will have $|M_{fi}| < \gamma$, and the levels will be sufficiently broadened for perturbation theory. This will occur if the allowed contact current is greater than device current which is limited by tunneling transmission $\mathcal{T}_{\text{device}}$. For a single level device such as the 0d-0d and 1d-$1d_{edge}$, the contact current can be related[63] to the broadening γ to give: $I_{contact} = 2q\gamma/\hbar$, from Section VII, where the factor 2 is due to spin. The device current is given by Fermi's golden rule:

$$I_{device} = 2q \times \frac{2\pi}{\hbar} \times \left|M_{fi}\right|^2 \times \frac{dN}{dE} \tag{6.11.1}$$

The density of states dN/dE can also be expressed as the inverse of the level spacing, ΔE, or level broadening, γ: $dN/dE = 1/\Delta E$ if $\Delta E < \gamma$, or $dN/dE = 1/\gamma$ if $\Delta E > \gamma$. Using this and setting $I_{device} < I_{contact}$ gives $|M_{fi}|^2 < (\gamma \times \Delta E)$ or $|M_{fi}|^2 < \gamma^2$, respectively. But in the first case $\Delta E < \gamma$ assures $|M_{fi}| < \gamma$. Therefore in both cases $|M_{fi}| < \gamma$ satisfies the level broadening requirements for Fermi's Golden Rule, and limits the permitted tunneling current.

Perturbation theory requires the matrix element to be less than the width of the broadening. We have shown that the levels are sufficiently broadened when the device current is limited by the tunneling junction but not limited by the contacts. The same restriction on tunnel junction current applies to the 2d-$2d_{face}$ case, since both the contact current and the device current are multiplied by the number of transverse y states and can be analyzed as many 1d-$1d_{edge}$ modes in parallel.

The tunneling matrix element, which is less than γ, is given by Eqn. (6.16.12): $\left|M_{fi}\right| = \frac{1}{\pi}\sqrt{E_{Z,i} \times E_{Z,f} \times \mathcal{T}_{\text{device}}}$. Solving for the maximum permitted tunneling transmission probability $\mathcal{T}_{\text{device}}$ in the general case:

$$\mathcal{T}_{\text{device}} < \frac{\pi^2 \gamma^2}{E_{Z,i} \times E_{Z,f}} \tag{6.11.2}$$

This perturbation requirement applies to the 0d-0d, 1d-$1d_{edge}$, and 2d-$2d_{face}$ cases.

Now specifically considering the 0d-0d case, the contact broadening is Eqn. (6.8.4), $\gamma = (1/\pi) \times E_{Z,i} \times \mathcal{T}_{\text{contact}}$. Inserting this into the maximum permitted tunneling transmission probability, Eqn. (6.11.3), with the E_Z confinement energies equal, the basic requirement $|M_{fi}| < \gamma$ implies:

$$\mathcal{T}_{\text{device}} < \mathcal{T}_{\text{contact}}^2 \tag{6.11.3}$$

which we regard as the condition for the validity of time-dependent perturbation theory for the 0d-0d case.

The contact broadening for a specific 1d-1d$_{edge}$ case will now be worked out:

We consider the 1d-1d$_{edge}$ case with contacts consisting of nanowires extending to infinity. The electrical contact broadening for 1d-1d edge overlapped nanowires, Figure 6.11(a), is controlled by the loss of carriers into the extended nanowires. Let the length of the overlap region be L_X. In each wire, the carrier will travel an average distance of L_X before escaping. At turn on, its momentum is given by $p_X=mv_X=\hbar k_X$ where $k_X=\pi/L_X$ in the ground state. Using $E_X=\hbar^2k_X^2/2m=\hbar^2\pi^2/2mL_X^2$, the average escape time is $\tau=L_X/v_X=h/4E_X$. Then the energy level broadening due to contacts is:

$$\gamma\equiv\hbar/\tau=(2/\pi)E_X=(2/\pi)\hbar^2\pi^2/2mL_X^2 \tag{6.11.4}$$

The coupling matrix element is the same as the 0d-0d case and is given by Eqn. (6.16.12): $\left|M_{fi}\right|=\frac{1}{\pi}\sqrt{E_{Z,i}\times E_{Z,f}\times\mathcal{T}_{device}}$. Requiring $|M_{fi}|<\gamma$ implies that the tunneling transmission factor $\mathcal{T}_{device}$ should not be too large:

$$\sqrt{\mathcal{T}_{device}}<\frac{2L_{Z,i}\times L_{Z,f}}{L_X^2} \tag{6.11.5}$$

This is the condition for perturbation theory to be valid for 1d-1d$_{edge}$ case contacted by extended wires. The same condition applies to the 2d-2d$_{face}$ case because it can be analyzed as many 1d-1d$_{edge}$ modes in parallel. This condition can be relaxed if the level broadening is dominated by a scattering mechanism such as electron-phonon scattering.

6.12 MAXIMUM CONDUCTANCE LIMIT

Since we have found a maximum permitted tunneling transmission, $\mathcal{T}_{device}$ for perturbation theory, this sets the maximum permitted current or conductance within our perturbation approach. For the 0d-0d case we already derived Eqn. (6.8.10), the conductance $G/(q^2/h)=2\pi^2\gamma/(k_bT)\times[T_{device}/(T_{contact})^2]$. The maximum permitted tunneling transmission is $[T_{device}/(T_{contact})^2]=1$, which leads to the maximum permitted conductance:

$$G_{0d-0d}<\frac{2q^2}{h}\times\pi^2\times\frac{\gamma}{k_bT} \tag{6.12.1}$$

which is unfortunately less than the conductance quantum for sharp thresholds, $\gamma<k_BT$.

For the 1d-1d$_{edge}$ case the expression for the 1d-1d$_{edge}$ conductance is Eqn. (6.10.2):

$$G_{1d-1d,edge}=2\frac{L_X}{\pi^2\hbar^2}E_{Z,i}\times E_{Z,f}\times\sqrt{\frac{m}{qV_{OL}}}\times\mathcal{T}_{device}\times\frac{q}{4k_bT}$$

In the most general damping model $\gamma>\left|M_{fi}\right|=\frac{1}{\pi}\sqrt{E_{Z,i}\times E_{Z,f}\times\mathcal{T}_{device}}$. Plugging $|M_{fi}|=\gamma$ into $G_{1d\text{-}1d,\,edge}$ and setting $qV_{OL}=\gamma$ to provide the peak permitted conductance gives:

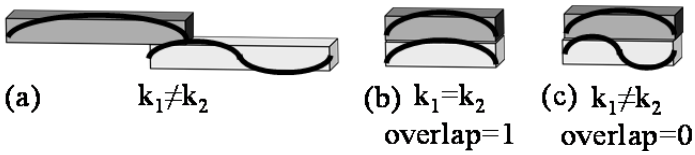


Figure 6.12: (a) When there is poor spatial overlap, at long wave-vectors near the turn-on threshold, the overlap integral is small and nonzero. (b) For perfect spatial overlap, the overlap integral between the same transverse k-vector is 1. (c) The overlap integral is zero for different transverse k-vectors.

$$G_{1d-1d,edge} < \frac{2q^2}{h} \times \sqrt{2}\pi^2 \times \gamma^{3/2} \times \frac{1}{4k_b T} \times \sqrt{\frac{2mL_X^2}{\pi^2\hbar^2}} \tag{6.12.2}$$

Now we specialize to damping due to extended nanowire contacts reaching to infinity. Inserting the broadening equation, Eqn. (6.11.4), $\gamma=(2/\pi)E_X$, and the explicit expression $E_X = \hbar^2\pi^2/2mL_X^2$ gives the maximum conductance:

$$G_{1d-1d,edge} < \frac{2q^2}{h} \times \frac{\pi^{3/2}}{2} \times \frac{\gamma}{k_B T} \tag{6.12.3}$$

which is a similar limit as the 0d-0d case.

The derivation for the 2d-2d$_{face}$ case maximum permitted conductance is similar:

From Eqn. (6.9.3):

$$G_{2d-2d,face} = \frac{qmA}{\pi^2\hbar^3} \times E_{Z,i} \times E_{Z,f} \times \mathcal{T}_{device} \times \frac{q}{4k_b T}$$

Where $A \equiv W \times L_X$ is the area of the overlap region between quantum wells. Once again, in the most general damping model $\gamma = |M_{fi}| = \frac{1}{\pi}\sqrt{E_{Z,i} \times E_{Z,f} \times \mathcal{T}_{device}}$. Plugging $|M_{fi}|=\gamma$ into $G_{2d\text{-}2d,\ face}$, we obtain the general maximum permitted conductance:

$$G_{2d-2d,face} < \frac{2q^2}{h} \times \frac{\pi^3}{2} \times \gamma^2 \times \frac{1}{4k_b T} \times \left(\frac{2mWL_X}{\hbar^2\pi^2}\right) \tag{6.12.4}$$

Now we specialize to damping owing to extended quantum well contacts reaching to infinity. Inserting the broadening equation, Eqn. (6.11.4), $\gamma=(2/\pi)E_X$, and the explicit expression $E_X = \hbar^2\pi^2/2mL_X^2$ gives the maximum conductance:

$$G_{2d-2d,face} < \frac{2q^2}{h} \times \frac{\pi^2}{4} \times \frac{\gamma}{k_b T} \times \frac{W}{L_X} \tag{6.12.5}$$

Thus we have obtained the maximum permitted perturbation conductance for the 0d-0d, 1d-1d, and 2d-2d cases, for both general damping and for end-wire damping models. Within the limits of perturbation theory all of the cases have a tradeoff between the broadening and the conductance.

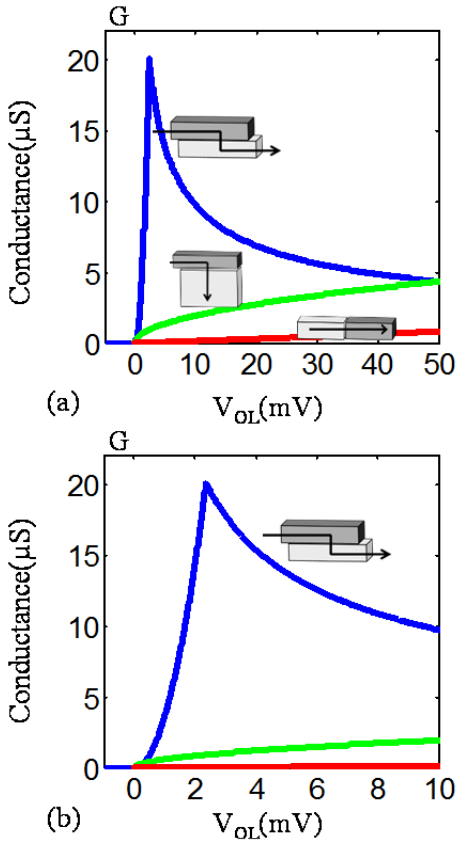


Figure 6.13: (a) The conductance curves for various 1d dimensionalities are plotted. Parameters were chosen at the limit of perturbation theory: Tunneling transmission probability is T_{device}=2.16%; damping due to propagation down the nano-wire is γ=2.34meV; nanowire thickness and corresponding quantum confinement energy is L_Z=8.7nm and E_Z=50meV respectively; and length of overlap region L_X=32nm. The effective mass was $0.1m_e$. (b) A close-up view.

6.13 SMEARING THE STEEP RESPONSE:

The finite overlap length L_X of the 1d-$1d_{edge}$, and 2d-$2d_{face}$ cases led to loss of carriers to the extended nano-wire/quantum well contacts, and the damping smeared the steep response. We can show the modification of the steep response in an entirely different way, by employing the exact wave functions at the ends of the wires. The wave-functions are illustrated in Figure 6.12(a) where long wavelengths approach zero at the ends of the wires, precisely where overlap is needed. These long wavelengths occur right at threshold, impairing the sharp turn-on.

On the other hand, shorter wavelengths as shown Figure 6.12(b) and (c) are either overlapping in Figure 6.12(b) or orthogonal in Figure 6.12(c). If there is a perfect overlap the states completely couple by tunneling.

The shape of the turn on due to wave functions going to zero at the wire ends can be modeled following the methods in the Appendix. The kronecker deltas in Eqn. (6.16.10) need to be replaced by the actual transverse overlap integral from Eqn. (6.16.5) and the sums in Eqn.

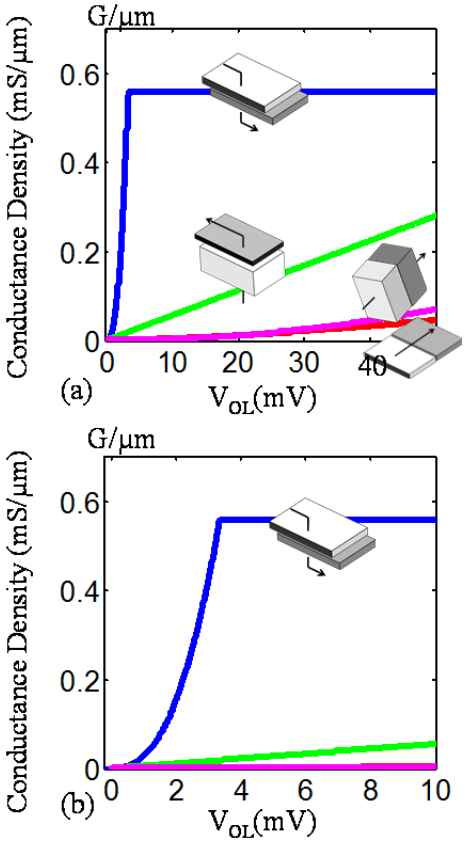


Figure 6.14: .(a) The conductance curves for various 2d and 3d dimensionalities are plotted. The parameters chosen were the same as in Figure 6.13, at the limit of perturbation theory: Tunneling transmission probability is $\mathcal{T}_{device}$=2.16%; damping due to propagation down the nano-wire is γ =2.34meV; nanowire thickness and corresponding quantum confinement energy is L_Z=8.7nm and E_Z=50meV respectively; and length of overlap region L_X=32nm. The effective mass was $0.1m_e$. (b) A close-up view.

(6.17.4) should include all the transverse states. Taking the limit near turn-on for long k-vectors, we get the following expressions for the conductivities:

$$G_{1d-1d,edge,\,turn\text{-}on} \approx \frac{2q^3}{h} \times \frac{\pi^5}{72} \times \frac{E_{Z,i}E_{Z,f}}{E_X{}^3} \times V_{OL}^2 \times \mathcal{T}_{device} \times \frac{q}{4k_bT} \tag{6.13.1}$$

$$G_{2d-2d,face,\,turn\text{-}on} \approx \left(\frac{2q^3}{h} \times \frac{\pi^5}{72} \times \frac{E_{Z,i}E_{Z,i}}{E_X{}^3} \times V_{OL}^2 \times \mathcal{T}_{device} \times \frac{q}{4k_bT} \right) \times \frac{14}{15} \times \frac{W}{\pi\hbar} \sqrt{\frac{qmV_{OL}}{2}} \tag{6.13.2}$$

where, W is the width of the quantum well. The turn-on conductance versus overlap control voltage V_{OL} can be seen in Figure 6.13 for the 1d-1d$_{edge}$, 1d-2d, and 1d-1d$_{point}$ cases. Similarly Figure 6.14 covers the 2d-2d$_{face}$, 2d-3d, 3d-3d and 2d-2d$_{edge}$ cases, with conductance per unit width plotted.

The effective broadening is the overlap voltage required to reach peak conductance for the 1d-1d$_{edge}$ and 2d-2d$_{face}$ cases that would otherwise have been infinitely sharp.

For the 1d-1d$_{edge}$ case, the voltage requirement is met when Eqn. (6.13.1) equals the on state conductance Eqn. (6.10.2): $G_{1d-1d,edge} = 2\frac{Lq}{\pi^2\hbar^2} E_{Z,i} \times E_{Z,f} \times \sqrt{\frac{m}{qV_{OL}}} \times \mathcal{T}_{device} \times \frac{q}{4k_bT}$. This gives an effective broadening:

$$\gamma \equiv qV_{OL} = \left(\frac{72\sqrt{2}}{\pi^5}\right)^{\frac{2}{3}} E_X = 0.48\, E_X \tag{6.13.3}$$

Likewise for the 2d-2d case the voltage requirement is met when Eqn. (6.13.2) equals the on state conductance Eqn. (6.9.3) $G_{2d-2d,face} = \frac{qmA}{\pi^2\hbar^3} \times E_{Z,i} \times E_{Z,f} \times \mathcal{T}_{device} \times \frac{q}{4k_bT}$. This gives an effective broadening:

$$\gamma \equiv qV_{OL} = \left(\frac{540}{7\pi^4}\right)^{\frac{2}{5}} E_X = 0.91\, E_X \tag{6.13.4}$$

In both cases $\gamma \sim E_X$ which is what we found in the simple escape time model given by Eqn. (6.11.4).

6.14 Density of States Broadening

In sections 6.1 to 6.10 we have assumed an ideal 1d, 2d or 3d density of states and ignored any level broadening. In practical devices the band edges will not have an ideal density of states that fall sharply to zero at the band edge, but rather there will be a band tail. This tail will be caused by any imperfections in the lattice, whether they are due to impurities or phonons. In optical measurements this results in the Urbach tail of the absorption spectrum. In silicon the optical absorption coefficient falls off as an exponential at the rate of 23mV/decade[25, 26]. If a similar tail exists in tunneling devices it could also pose a limit on the achievable sub-threshold slope and on the level broadening, γ.

6.15 Conclusions

Since there were many geometries, and many different cases covered here, we provide a global Table I that covers all the cases considered in this paper.

Dimensionality significantly affects the low voltage turn on characteristics of a tunneling device, including Backward Diodes and tunneling Field Effect Transistors. The ideal tunneling transistor would have step function turn on characteristic. Fortunately, a 2d-2d$_{face}$ overlapped tunneling junction is very close to this. In practice, various effects such as nonuniformities, dopants, phonons, series resistance, level broadening, poor wavefunction overlap, will prevent us from observing an ideal 2d density of states step function turn on. In spite of non-idealities the 2d-2d$_{face}$ overlapped junction is expected to bring us closer to a step function response.

Furthermore quantum confinement on either side of a tunneling barrier can significantly boost the on-state conductance.

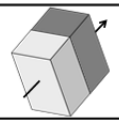
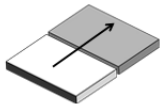
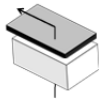
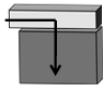
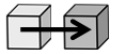
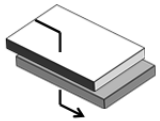
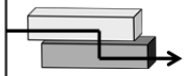
Case	Picture	Current	Conductance, G	Maximum G for pert. theory to be valid	Maximum G for end contacts $\gamma = (2/\hbar)E_X$
1d-1d		$\frac{2q^2}{h} \times V_{OL} \times \mathcal{T}_{device}$	$\frac{2q^2}{h} \times V_{OL} \times \mathcal{T}_{device} \times \frac{q}{4k_bT}$	N/A	N/A
3d-3d		$\frac{Am^*}{4\pi\hbar^2} \times \frac{qV_{OL}}{2} \times \frac{2q^2}{h} V_{OL} \times \mathcal{T}_{device}$	$\frac{Am^*}{4\pi\hbar^2} \times \frac{qV_{OL}}{2} \times \frac{2q^2}{h} V_{OL} \times \mathcal{T}_{device} \times \frac{q}{4k_bT}$	N/A	N/A
2d-2d$_{edge}$		$\frac{2L_X\sqrt{qm^*V_{OL}}}{3\pi\hbar} \times \frac{2q^2}{h} \times V_{OL} \times \mathcal{T}_{device}$	$\frac{2L_X\sqrt{qm^*V_{OL}}}{3\pi\hbar} \times \frac{2q^2}{h} \times V_{OL} \times \mathcal{T}_{device} \times \frac{q}{4k_bT}$	N/A	N/A
0d-1d		$\frac{2q}{h} \times E_Z \times \mathcal{T}_{device}$	$\frac{2q}{h} \times E_Z \times \mathcal{T}_{device} \times \frac{q}{4k_bT}$	N/A	N/A
2d-3d		$\frac{Am}{2\pi\hbar^2} \times \frac{qV_{OL}}{2} \times \frac{4q}{h} \times E_Z \times \mathcal{T}_{device}$	$\frac{Am}{2\pi\hbar^2} \times \frac{qV_{OL}}{2} \times \frac{4q}{h} \times E_Z \times \mathcal{T}_{device} \times \frac{q}{4k_bT}$	N/A	N/A
1d-2d		$\frac{L_X}{\pi\hbar} \times \sqrt{qm^*V_{OL}} \times \frac{4q}{h} \times E_Z \times \mathcal{T}_{device}$	$\frac{L_X}{\pi\hbar} \times \sqrt{qm^*V_{OL}} \times \frac{4q}{h} \times E_Z \times \mathcal{T}_{device} \times \frac{q}{4k_bT}$	N/A	N/A
0d-0d		$\frac{4q}{\hbar} E_{Z,i} \times \frac{\mathcal{T}_{device}}{\mathcal{T}_{contact}}$	$\frac{4q}{\hbar} E_{Z,i} \times \frac{\mathcal{T}_{device}}{\mathcal{T}_{contact}} \times \frac{q}{4k_bT}$	$\frac{2q^2}{h} \times \pi^2 \times \frac{\gamma}{k_bT}$	$\frac{2q^2}{h} \times \pi^2 \times \frac{\gamma}{k_bT}$
2d-2d$_{face}$		$\frac{qmA}{\pi^2\hbar^3} \times E_{Z,i} \times E_{Z,f} \times \mathcal{T}_{device}$	$\frac{qmA}{\pi^2\hbar^3} \times E_{Z,i} \times E_{Z,f} \times \mathcal{T}_{device} \times \frac{q}{4k_bT}$	$\frac{2q^2}{h} \times \frac{\pi^3}{2} \times \gamma^2 \times \frac{1}{4k_bT} \times \frac{2mA}{\hbar^2\pi^2}$	$\frac{2q^2}{h} \times \frac{\pi^2}{4} \times \frac{W}{L_X} \times \frac{\gamma}{k_bT}$
1d-1d$_{edge}$		$2\frac{qL_X}{\pi^2\hbar^2} E_{Z,i} \times E_{Z,f} \times \sqrt{\frac{m}{qV_{OL}}} \times \mathcal{T}_{device}$	$2\frac{qL_X}{\pi^2\hbar^2} E_{Z,i} \times E_{Z,f} \times \sqrt{\frac{m}{qV_{OL}}} \times \mathcal{T}_{device} \times \frac{q}{4k_bT}$	$\frac{2q^2}{h} \times \sqrt{2}\pi^2 \times \gamma^{3/2} \times \frac{1}{4k_bT} \times \sqrt{\frac{2mL_X^2}{\pi^2\hbar^2}}$	$\frac{2q^2}{h} \times \frac{2\pi^{3/2}}{4} \times \frac{\gamma}{k_bT}$

Table I. Comparison of the nine different dimensionalities

6.16 APPENDIX A: TRANSFER-MATRIX ELEMENT DERIVATION

In our derivation of the tunnel matrix element by the transfer Hamiltonian method we will consider 3d-3d case. The method for the other reduced dimensionality cases is very similar and we will note some of the changes that would be necessary for those cases as we go through the derivation.

First we consider a simple Type III junction band diagram as shown in Figure 6.15. The total Hamiltonian H, is illustrated in Figure 6.15(a). The incomplete initial Hamiltonian H_i, on the left is in Figure 6.15(b), and the incomplete final state Hamiltonian H_f on the right is in Figure 6.15(c). For the cases in Figure 6.15(b)&(c), the incomplete Hamiltonians lead to their own stationary Schrodinger's equations: $H_i|\Psi_i\rangle=E_i|\Psi_i\rangle$ and $H_f|\Psi_f\rangle=E_f|\Psi_f\rangle$ respectively. The subscript 'i' represents the initial electron in the valence band and the subscript 'f' represents the final electron in the conduction band.

In the true full Hamiltonian, H, a valence band electron on the left decays exponentially into the barrier, and tunnels to the conduction band on the right. The perturbation Hamiltonian with respect to the starting Hamiltonian is therefore $H'=H-H_i$. The Fermi's Golden Rule transition rate for an electron in the valence band on the left, tunneling to the conduction band on the right, is.

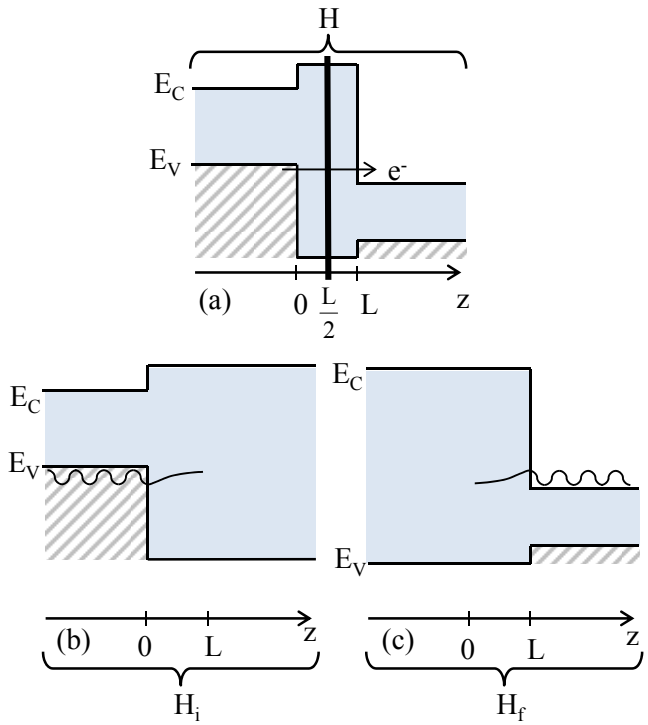


Figure 6.15: (a) The exact total Hamiltonian *H*. (b) The incomplete Hamiltonian H_i whose eigenstate represents the initial valence electron of energy E_i. (c); The incomplete Hamiltonian H_f whose eigenstate represents the final conduction band electron of energy E_f.

$$R_{if} = \frac{2\pi}{\hbar}\left|\langle \psi_f \mid H' \mid \psi_i \rangle\right|^2 \frac{dN}{dE} = \frac{2\pi}{\hbar}\left|\langle \psi_f \mid H - H_i \mid \psi_i \rangle\right|^2 \frac{dN}{dE} = \frac{2\pi}{\hbar}\left|\langle \psi_f \mid H - E_i \mid \psi_i \rangle\right|^2 \frac{dN}{dE} \quad (6.16.1)$$

where we used the fact that $H_i|\Psi_i\rangle = E_i|\Psi_i\rangle$, and dN/dE represents the density of final states.

The exact Hamiltonian, in Figure 6.15(a) naturally divides into three regions. For $z<0$ the system resembles H_i, whose eigenstates are in the valence band on the left. For $0<z<L$, there is a barrier which the electron must tunnel through, and for $z>L$ the system resembles H_f with eigenstates in the conduction band on the right. Ψ_i is a free particle in the valence band and the exponential decay can be modeled by the WKB approximation. For convenience we segregate the problem into halves, picking a surface somewhere in the barrier so that we can divide the junction into a left half and a right half. For simplicity we choose the dividing plane to be at L/2 as shown in Figure 6.15(a).

Since $(H_i - E_i)|\Psi_i\rangle = 0$ everywhere, and $H \equiv H_i$ in the left half space, then $(H - E_i)|\Psi_i\rangle = 0$, in the left half-space; $z<L/2$. Likewise, since $(H_f - E_f)|\Psi_f\rangle = 0$ everywhere, and $H \equiv H_f$ in the right half-space, then, $(H - E_f)|\Psi_f\rangle = 0$, in the right half-space; $z>L/2$.

Following refs. [57] &[58], the matrix element, $M_{fi} = \int_{-\infty}^{\infty} d^3r\, \psi_f^* (H - E_i)\psi_i$ can be simplified by recognizing that the integral is certainly zero for $z < L/2$ and by subtracting $0 = [\psi_i^* (H - E_f)\psi_f]^*$ for $z > L/2$. Further simplification arises when we express the Hamiltonian in the standard format:

$$H = -\frac{\hbar^2 \nabla^2}{2m} + V(r) \quad (6.16.2)$$

where V(r) describes the entire potential of the junction. By substituting this into:

$$M_{fi} = \int_{Z>L/2}^{\infty} d^3r \left[\psi_f^* (H - E_i)\psi_i - \psi_i (H - E_f)\psi_f^*\right] \quad (6.16.3)$$

and using both energy conservation, $E_i = E_f$, and the cancellation of terms involving V(r), we will be left with:

$$\begin{aligned} M_{fi} &= \frac{-\hbar^2}{2m}\int_{Z>L/2} d^3r \times \left(\psi_f^* \nabla^2 \psi_i - \psi_i \nabla^2 \psi_f^*\right) \\ &= \frac{-\hbar^2}{2m}\int_{Z>L/2} d^3r \times \nabla \cdot \left(\psi_f^* \nabla \psi_i - \psi_i \nabla \psi_f^*\right) \end{aligned} \quad (6.16.4)$$

Now we use Gauss's law to express the matrix element as:

$$M_{fi} = \hbar i \int_{Z=L/2} \vec{G}_{fi} \cdot d\vec{S} \text{ where} \quad (6.16.5)$$

$$\text{with } \vec{G}_{fi} \equiv \frac{i\hbar}{2m}\left(\psi_f^* \nabla \psi_i - \psi_i \nabla \psi_f^*\right) \quad (6.16.6)$$

The matrix element is now expressed as a surface integral of $\vec{G}_{fi}$ which is nonzero only at the z=L/2 surface.

To determine the tunneling matrix element (A.5) in our case of 3d-3d bulk tunneling we must first write down Ψ_i and Ψ_f in order to evaluate G_{fi}. Within the effective mass approximation, we can use the WKB approximation to write down the wave functions. We neglect the underlying Bloch functions, but for a more complete treatment see ref. [59]. We also assume that most of the probability density is outside of the barrier region and so the barrier region can be neglected when calculating the normalization constant. The normalized WKB approximation becomes:

$$\psi_i = \sqrt{\frac{2k_{Z,i}}{L_X L_Y L_{Z,i}}} \times \exp\left(ik_{X,i}x + ik_{Y,i}y\right) \times \frac{1}{\sqrt{k_Z(z)}} \times \begin{cases} \sin\left(\int_Z^0 k(z') \times dz' + \frac{\pi}{4}\right), & z < 0 \\ \frac{1}{2}\exp\left(-\int_0^Z k(z') \times dz'\right), & z \geq 0 \end{cases} \tag{6.16.7a}$$

$$\psi_f = \sqrt{\frac{2k_{Z,f}}{L_X L_Y L_{Z,f}}} \times \exp\left(ik_{X,f}x + ik_{Y,f}y\right) \times \frac{1}{\sqrt{k_Z(z)}} \times \begin{cases} \sin\left(\int_L^Z k(z') \times dz' + \frac{\pi}{4}\right), & z > L \\ \frac{1}{2}\exp\left(-\int_Z^L k(z') \times dz'\right), & z \leq L \end{cases} \tag{6.16.7b}$$

In these equations $k_{\alpha,i}$ and $k_{\alpha,f}$ are the α-component of the k-vector in the initial and final states respectively. $k_Z(z)$ is the spatially dependent value of k_Z that varies within the barrier. L_X, L_Y, $L_{Z,i}$, and $L_{Z,f}$ are the dimensions of the device as shown in Figure 6.4(a). $L_{Z,i}$ represents the length of the left half of the device for $z < 0$. $L_{Z,f}$ represents the length of the right half of the device for $z>L$. Plugging these wavefunctions into $\vec{G}_{fi}$ and evaluating it at $z=L/2$ gives:

$$G_{fi,\hat{Z}} = -\sqrt{\frac{k_{Z,f}k_{Z,i}}{L_{Z,f}L_{Z,i}}} \frac{i\hbar}{2mL_X L_Y} \exp\left(i\Delta k_X x + i\Delta k_Y y\right) \times \exp\left(-\int_0^L k_Z dz\right) \tag{6.16.8}$$

where $\Delta k_X = (k_{X,i} - k_{X,f})$ and $\Delta k_Y = (k_{Y,i} - k_{Y,f})$. Using this and evaluating the expression for the matrix element we get:

$$M_{fi} = \frac{\hbar^2}{2m}\sqrt{\frac{k_{Z,f}k_{Z,i}}{L_{Z,f}L_{Z,i}}} \times \exp\left(-\int_0^L k_Z dz\right) \times \delta_{k_{X,i},k_{X,f}} \delta_{k_{Y,i},k_{Y,f}} \tag{6.16.9}$$

The kronecker deltas represent the conservation of transverse momentum and show that the conservation is a natural result of calculating the matrix element. For the case of incomplete conservation of momentum, the kronecker deltas will be replaced by the actual surface integral in Eqn. (6.16.4). At this point we desire to replace $\exp\left(-2\int_0^L k_Z dz\right)$ with $\mathcal{T}_{\text{device}}$. But we redefine $\mathcal{T}_{\text{device}}$ to be a phenomenological factor that includes both the WKB exponential and the effect of the underlying Bloch functions. Thus the matrix element is given by:

$$M_{fi} = \frac{\hbar^2}{2m}\sqrt{\frac{k_{Z,f}k_{Z,i}}{L_{Z,f}L_{Z,i}}} \times \sqrt{\mathcal{T}_{\text{device}}} \times \delta_{k_{X,i},k_{X,f}} \delta_{k_{Y,i},k_{Y,f}} \tag{6.16.10}$$

Interestingly, this expression is also valid for all of the reduced dimensionality cases, we just need to sum over fewer k-states.

For the reduced dimensionality cases we can use $k_Z=\pi/L_Z$ and $E_Z = \hbar^2 k_Z^2 / 2m^*$ to further simplify the matrix element. For 0d-1d we get the following matrix element:

$$M_{fi,0d-1d} = \sqrt{\frac{E_{Z,i}}{\pi} \times \left(\frac{\hbar^2}{2m} \frac{k_{Z,f}}{L_{Z,f}} \right) \times \mathcal{T}_{device}} \tag{6.16.11}$$

For 0d-0d both sides of the junction are confined which gives:

$$M_{fi,0d-0d} = \frac{1}{\pi} \sqrt{E_{Z,i} \times E_{Z,f} \times \mathcal{T}_{device}} \tag{6.16.12}$$

6.17 Appendix B: Using the Transfer Matrix Element to Derive Current

In Appendix A we found the matrix element that can be used with Fermi's golden rule. Using this, we can now find the current in any of the different cases. To aid in correctly counting the number of states, we use the delta function version of Fermi's golden rule. The transition rate between two states is:

$$R_{if} = \frac{2\pi}{\hbar} \left| \langle \psi_f \mid H - E_i \mid \psi_i \rangle \right|^2 \delta(E_i - E_f) \tag{6.17.1}$$

We convert the transition rate to a tunneling current by multiplying the rate by the electron charge, summing over initial and final states, and multiplying by the Fermi-Dirac occupation probabilities.

$$J_{Tunnel} = 2q \sum_{k_i, k_f} R_{if} f_V (1 - f_C) - R_{fi} f_C (1 - f_V) \tag{6.17.2a}$$

$$= 2q \sum_{k_i, k_f} R_{fi} (f_C - f_V) \tag{6.17.2b}$$

$$= \frac{4\pi q}{\hbar} \sum_{k_i, k_f} \left| M_{fi} \right|^2 \delta(E_C - E_V)(f_C - f_V) \tag{6.17.2c}$$

Where

$$f_V = \frac{1}{\exp\left[(E_i - F_p)/k_b T\right] + 1} \tag{6.17.3a}$$

$$f_C = \frac{1}{\exp\left[(E_f - F_n)/k_b T\right] + 1} \tag{6.17.3b}$$

F_n and F_p are the quasi Fermi levels for electrons and holes respectively.

Plugging the matrix element Eqn. 6.16.10 into Eqn. 6.17.2c for tunneling current gives:

$$I_{Tunnel} = \frac{\pi q \hbar^3}{m^2} \sum_{k_i,k_f} \frac{k_{Z,i} k_{Z,f}}{L_{Z,i} L_{Z,f}} (f_C - f_V) \times \mathcal{T}_{device} \times \delta_{k_{X,i},k_{X,f}} \delta_{k_{Y,i},k_{Y,f}} \delta(E_i - E_f) \quad (6.17.4a)$$

$$I_{Tunnel} = \frac{8q}{h} \sum_{k_i,k_f} \left(\frac{\hbar^2 k_{Z,i}}{2m} \frac{\pi}{L_{Z,i}} \right) \times \left(\frac{\hbar^2 k_{Z,f}}{2m} \frac{\pi}{L_{Z,f}} \right) \times (f_C - f_V) \times \mathcal{T}_{device} \times \delta_{k_{X,i},k_{X,f}} \delta_{k_{Y,i},k_{Y,f}} \delta(E_i - E_f) \quad (6.17.4b)$$

Interestingly, this expression is also valid for all of the reduced dimensionality cases, we just need to sum over fewer k-states.

For 3d to 3d bulk we break the sums up into a transverse (k_t) and z component (k_Z) and then we convert the sums over k_Z and k_t to integrals over the z-component of energy (E_Z) and transverse energy (E_t) respectively. We can then convert the integrals over E_Z to integrals over total energy by a change of variables. Subsequently evaluating the delta functions gives us:

$$I_{Tunnel} = \frac{qmA}{2\pi^2 \hbar^3} \int_0^{V_{OL}} dE_i \int_0^{\min(E, V_{OL} - E)} dE_t (f_C - f_V) \times \mathcal{T}_{device} \quad (6.17.5)$$

Here we take the zero of energy to be at the conduction band edge on the n-side. The transverse energy can be no more than the total energy on either side of the junction. For reduced dimensionalities we will be summing over fewer k-states and so there may be only one or even no integrals.

Now we can set $(f_C - f_V) \approx qV_{SD}/(4k_BT)$ by assuming small biases less than k_BT. Finally, we evaluate the integrals and divide by V_{SD} to recover Eqn. (6.3.3)

$$G_{3d-3d} = \frac{1}{2} \left(\frac{Am^*}{2\pi\hbar^2} \times \frac{qV_{OL}}{2} \right) \times \frac{2q^2}{h} V_{OL} \times \mathcal{T}_{device} \times \frac{q}{4k_bT} \quad (6.17.6)$$

Thus we have finally recovered the equation for 3d-3d bulk tunneling current with small biases.

Next, we consider the 2d-3d case to demonstrate the general applicability of Eqn. (6.17.4). In this case we sum over the transverse states (k_t) and only the final k_Z states. After converting the sums to energy integrals and evaluating the delta function we get:

$$I_{2d-3d} = \frac{Am}{2\pi\hbar^2} \times \frac{4q}{h} \times \int_0^{\min(E, V_{OL} - E)} dE_t (f_C - f_V) \times \mathcal{T}_{device} \quad (6.17.7)$$

Taking the small bias limit and dividing by V_{SD} we recover Eqn. (6.6.2)

$$G_{2d-3d} = \left(\frac{Am}{2\pi\hbar^2} \times \frac{qV_{OL}}{2} \right) \times \left(\frac{4q}{h} \times E_Z \times \mathcal{T} \right) \times \left(\frac{q}{4k_bT} \right) \quad (6.17.8)$$

Similarly, we can derive the current for any of the cases using Eqn. (6.17.4).

Chapter 7: Non-Equilibrium Green's Function Modeling of Dimensionality Effects

7.1 INTRODUCTION

Now that we have analytically demonstrated the importance of considering the dimensionality of a PN junction, we will take a different approach to the problem and numerically model the tunneling junctions and verify the key effects. The non-equilibrium greens function (NEGF) method allows us to accurately model all of the quantum effects and conservation laws.

In this chapter we will first explain the theory required to use a NEGF model starting with a one dimensional two band k.p model and then a two dimensional 8 band k.p model. Finally we will use the NEGF simulation to analyze a 1d-1d$_{edge}$ junction and show the key features of the junction. The basic NEGF theory follows Datta's book [63], but several extensions need to be made to account for the multiple bands[66-68]. The use of k.p theory follows from the work by Ganapathi et all [55].

7.2 1D 2 BAND K.P NEGF MODEL

In order to demonstrate the concepts behind the NEGF simulation we will first develop a one dimensional two band k.p model. The first step is to create a Hamiltonian to describe the system. To do this we first start with the following bulk two band k.p Hamiltonian[69]:

$$[H] = \begin{pmatrix} \frac{\hbar^2 k^2}{2m_0} + E_G & \frac{\hbar}{m_0} k \cdot p \\ \frac{\hbar}{m_0} k \cdot p & \frac{\hbar^2 k^2}{2m_0} \end{pmatrix} \tag{7.2.1}$$

This defines the band structure, but only describes a bulk material. In order to work in real space we need to convert k to a spatial derivative:

$$\vec{k} \rightarrow -i\vec{\nabla} \tag{7.2.2}$$

Thus we have:

$$[H]\Psi_n = \begin{pmatrix} -\frac{\hbar^2 \vec{\nabla}^2}{2m_0} + E_G + U_n & -i\frac{\hbar}{m_0} \vec{\nabla} \cdot \vec{p} \\ -i\frac{\hbar}{m_0} \vec{\nabla} \cdot \vec{p} & -\frac{\hbar^2 \vec{\nabla}^2}{2m_0} + E_G + U_n \end{pmatrix} \begin{pmatrix} \psi_{a,n} \\ \psi_{b,n} \end{pmatrix} \tag{7.2.3}$$

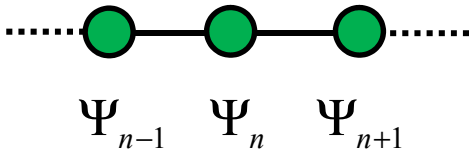


Figure 7.1: Consider a chain of 1d atoms

Now the wavefunction, Ψ_n, is dependent on the position, n. We also added a spatially varying potential U_n At this point we need to define a lattice on which the wavefunction is defined as shown in Figure 7.1. Each lattice point has two orbitals for each band. The spatial derivatives can now be converted to discrete derivatives:

$$k\psi_{a,n} = -i\frac{\partial}{\partial x}\psi_{a,n} = \frac{-i}{2a}(\psi_{a,n+1} - \psi_{a,n-1}) \tag{7.2.4}$$

$$k^2\psi_{a,n} = -\frac{\partial^2}{\partial x^2}\psi_{a,n} = \frac{-1}{a^2}(\psi_{a,n+1} - 2\psi_{a,n} + \psi_{a,n-1}) \tag{7.2.5}$$

Plugging these back into the Hamiltonian gives us a recursive equation linking the different wavefunctions. To simplify the equation we can define the following two constants:

$$t_0 = \frac{\hbar^2}{2m_0a^2} \tag{7.2.6}$$

$$t' = \frac{\hbar p}{2m_0a} \tag{7.2.7}$$

Using these, the Hamiltonian for the first three atoms is:

$$[H]\Psi_n = \left(\begin{array}{cc|cc|cc|c} E_G+U_1+2t_0 & 0 & -t_0 & -it' & 0 & 0 & \\ 0 & U_1+2t_0 & -it' & -t_0 & 0 & 0 & \\ \hline -t_0 & it' & E_G+U_2+2t_0 & 0 & -t_0 & -it' & \\ it' & -t_0 & 0 & U_2+2t_0 & -it' & -t_0 & \cdots \\ \hline 0 & 0 & -t_0 & it' & E_G+U_3+2t_0 & 0 & \\ 0 & 0 & it' & -t_0 & 0 & U_3+2t_0 & \\ & & \cdots & & & & \cdots \end{array}\right)\left(\begin{array}{c} \psi_{a,1} \\ \psi_{b,1} \\ \psi_{a,2} \\ \psi_{b,2} \\ \psi_{a,3} \\ \psi_{b,3} \\ \cdots \end{array}\right) \tag{7.2.8}$$

We can use this Hamiltonian to plot the band structure by assuming the potential is zero. To find the band structure, we should define:

$$[H_0] = \begin{pmatrix} E_G+U_1+2t_0 & 0 \\ 0 & U_1+2t_0 \end{pmatrix} \tag{7.2.9}$$

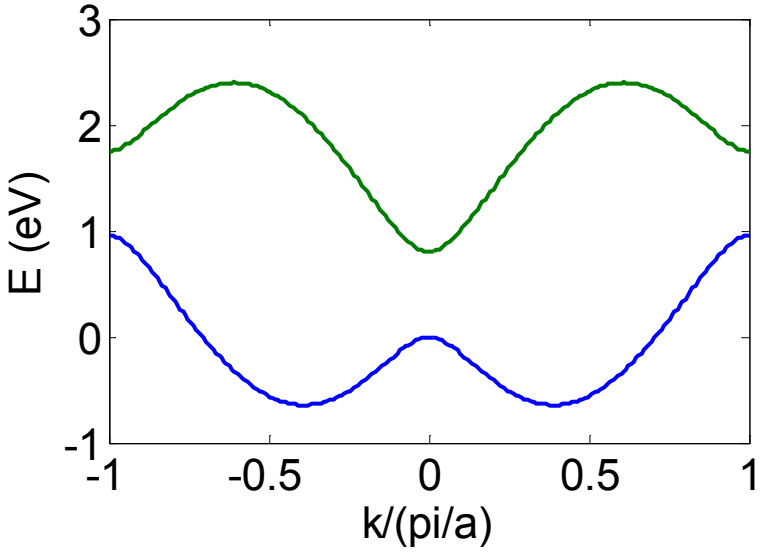


Figure 7.2: The band structure using nearest neighbor coupling is incorrect at large k values

$$[\tau] = \begin{pmatrix} -t_0 & it' \\ it' & -t_0 \end{pmatrix} \tag{7.2.10}$$

The bandstructure can be found by assuming a periodic wavefunction and finding the eigenvalues of the following equation[63]:

$$[H_k] = [H_0] + [\tau] \cdot \exp(ika) + [\tau]^+ \cdot \exp(-ika) \tag{7.2.11}$$

This gives the bandstructure shown in Figure 7.2. Near k=0 the band structure is correct. However, the band structure is very wrong at large k values and there are states in the band gap. This is because our model is based on a k.p Hamiltonian which only designed to be accurate in the zone center. This means that we need to somehow modify our Hamiltonian to eliminate the spurious states that exist in the band gap. The way to do this is to refine one of the derivatives to use next nearest neighbors as follows:

$$k^2 \psi_{a,n} = -\frac{\partial^2}{\partial x^2} \psi_{a,n} = \frac{-1}{4a^2} (\psi_{a,n+2} - 2\psi_{a,n} + \psi_{a,n-2}) \tag{7.2.12}$$

This gives the following Hamiltonian for the first three atoms:

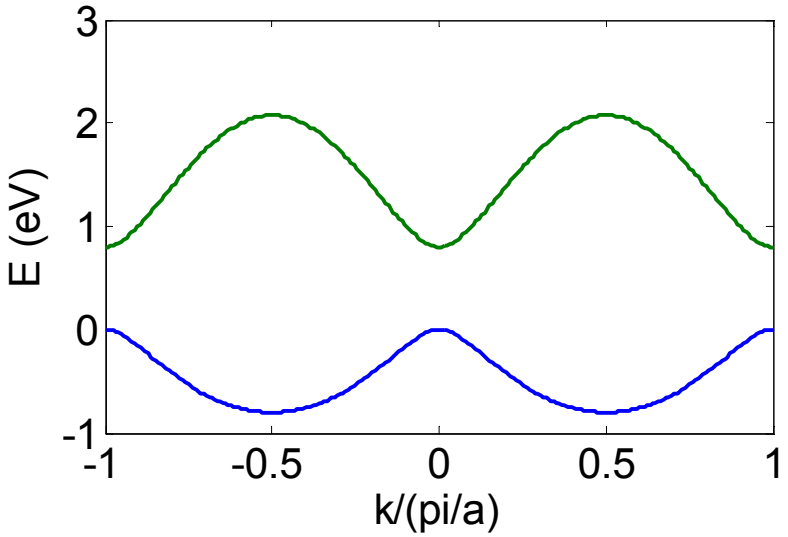


Figure 7.3: The band structure using next nearest neighbor coupling eliminates all spurious states in the band gap.

$$[H]=\left(\begin{array}{cc|cc|cc|c} E_G+U_1+\frac{t_0}{2} & 0 & 0 & -it' & -\frac{t_0}{4} & 0 & \\ 0 & U_1+\frac{t_0}{2} & -it' & 0 & 0 & -\frac{t_0}{4} & \\ \hline 0 & it' & E_G+U_2+\frac{t_0}{2} & 0 & 0 & -it' & \\ it' & 0 & 0 & U_2+\frac{t_0}{2} & -it' & -t_0 & \cdots \\ \hline -\frac{t_0}{4} & 0 & 0 & it' & E_G+U_3+\frac{t_0}{2} & 0 & \\ 0 & -\frac{t_0}{4} & it' & 0 & 0 & U_3+\frac{t_0}{2} & \\ & & \cdots & & & & \cdots \end{array}\right) \quad (7.2.13)$$

Using this Hamiltonian gives the band structure shown in Figure 7.3. All the spurious solutions have been eliminated. However, by requiring next nearest neighbor coupling we have effectively created a four band basis.

Now that we have a Hamiltonian that can describe our system, our next step is to find the electron density. The electron density can be represented by a sum of delta functions:

$$D(E)=\sum_{\alpha}\delta(E-\varepsilon_{\alpha}) \quad (7.2.14)$$

Next we can express the delta function as a limit:

$$2\pi\times\delta(E-\varepsilon_{\alpha})=\left.\frac{2\eta}{(E-\varepsilon_{\alpha})^2}\right|_{\eta\to 0^+}=i\left[\frac{1}{E-\varepsilon_{\alpha}+i\eta}-\frac{1}{E-\varepsilon_{\alpha}-i\eta}\right]_{\eta\to 0^+} \quad (7.2.15)$$

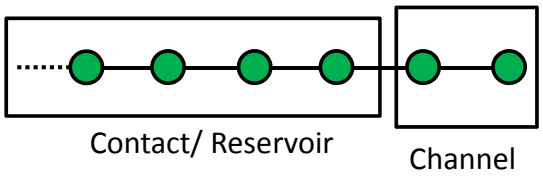


Figure 7.4: A larger system can be divided in to contact part and device or channel part

Next we can convert this into a matrix equation by replacing E with [H]. The validity of this substitution can be seen by considering the eigenstates of [H] as the basis. We now have:

$$[A] = i\left[G(E) - G^{+}(E)\right] \tag{7.2.16}$$

$$[G(E)] = \frac{1}{[H] - \varepsilon_{\alpha} + i\eta} \tag{7.2.17}$$

Thus we have now defined the Greens function for the system.

The next step is to find the greens function if we have a contact connected to our device. To analyze this consider a larger system that is divided into two parts, a channel and a contact/reservoir as shown in Figure 7.4. Next we need to accordingly divide our Hamiltonian:

Channel Hamiltonian (Size d X d) ← Small

$$[\overline{H}] = \begin{bmatrix} H & \tau \\ \tau^{+} & H_R \end{bmatrix}$$

τ ← Coupling between channel and contact (Size d X R)

H_R ← Contact Hamiltonian (Size R X R) ← Huge!!

(7.2.18)

Now we evaluate the Green's function:

$$[\overline{G}] = \left[(E + i0^{+})I - \overline{H}\right]^{-1} = \begin{bmatrix} (E + i0^{+})I - H & -\tau \\ -\tau^{+} & (E + i0^{+})I - H_R \end{bmatrix}^{-1}$$
$$= \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} G & G_{dR} \\ G_{Rd} & G_{RR} \end{bmatrix} \tag{7.2.19}$$

We are only interested in the electron density in the channel and so we only need to find the top left part of the matrix, [G]. Using the properties of matrix inversion we now have:

$$[G] = [A - BD^{-1}C]^{-1} = [(E + i0^{+})I - H - \Sigma]^{-1} \tag{7.2.20}$$

where

$$[\Sigma] = [\tau G_R \tau^{+}], [G_R] = \left[(E + i0^{+})I - H_R\right]^{-1} \tag{7.2.21}$$

By doing this we have abstracted the effect of the large contact into a self-energy term, [Σ]. If we had some way of finding [Σ], calculating [G] is easy.

The key to finding the self-energy is to assume an infinite contact where every atom is identical. In this case, the green's function at each atom will be identical. This means that we can set [G_R]=[G] in Eqn (7.2.21) and solve for [G] self consistently in Eqn (7.2.20) [68]. Thus we now have the electron density for an infinite contact.

Finally, to find the current we need to weight the electron density from the left contact by f1, weight he electron density from the right contact by f2 and evaluate:

$$I = \frac{d}{dt}\psi^{+}\psi \tag{7.2.22}$$

Using

$$i\hbar\frac{d\{\psi\}}{dt} = [H]\{\psi\} \tag{7.2.23}$$

The details of how to find the current are worked out nicely in Datta's book [63]. Thus we have completed the two band k.p NEGF model.

7.3 2D 8 BAND K.P NEGF MODEL

In order to model the dimensionality effects a two dimensional simulation is needed. However, when moving to two dimensions a simple two band model cannot reproduce the two dimensional band structure. Consequently we need to move to a more realistic band structure and use a 4 band or and 8 band k.p model. In our case we need to include the spin orbit coupling and use an 8 band model because we need to have the correct density of states to model the dimensionality effects. The 8 band k.p model is defined by the following equations [67, 70, 71]:

$$[H] = \begin{bmatrix} H_{\text{int}} & 0 \\ 0 & H_{\text{int}} \end{bmatrix} + [H_{S.O.}] \tag{7.3.1}$$

$$[H_{\text{int}}] = \begin{bmatrix} E_C + \frac{\hbar^2 k^2}{2m_o} + A'k^2 & +iP_o k_x & +iP_o k_y & +iP_o k_z \\ -iP_o k_x & \begin{matrix} E_V + \hbar^2 k^2/2m_o \\ +M(k_y^2 + k_z^2) + L'k_x^2 \end{matrix} & N'k_x k_y & N'k_x k_z \\ -iP_o k_y & N'k_x k_y & \begin{matrix} E_V + \hbar^2 k^2/2m_o \\ +M(k_x^2 + k_z^2) + L'k_y^2 \end{matrix} & N'k_y k_z \\ -iP_o k_z & N'k_x k_z & N'k_y k_z & \begin{matrix} E_V + \hbar^2 k^2/2m_o \\ +M(k_x^2 + k_y^2) + L'k_z^2 \end{matrix} \end{bmatrix} \tag{7.3.2}$$

$$[H_{S.O.}]=\frac{\Delta}{3}\begin{bmatrix}0&0&0&0&0&0&0&0\\0&-1&-i&0&0&0&0&1\\0&i&-1&0&0&0&0&-i\\0&0&0&-1&0&-1&i&0\\0&0&0&0&0&0&0&0\\0&0&0&-1&0&-1&i&0\\0&0&0&-i&0&-i&-1&0\\0&1&i&0&0&0&0&-1\end{bmatrix} \tag{7.3.3}$$

In this model we neglected Kane's B parameter which effectively means that we have assumed inversion symmetry and ignored the Dresselhaus effect.

P_o is the momentum matrix element and is given by:

$$P_o=-i\frac{\hbar}{m_0}\langle s\,|\,p_X\,|\,x\rangle \tag{7.3.4}$$

Given this we can define an energy E_P:

$$E_P=\frac{2m_0}{\hbar^2}P_0^2 \tag{7.3.5}$$

The constants L', M and N' can be defined in terms of the Luttinger parameters:

$$L'=-\frac{\hbar^2}{2m_0}\left(\gamma_1^L+4\gamma_2^L+1\right)+\frac{P^2}{E_G} \tag{7.3.6}$$

$$M=-\frac{\hbar^2}{2m_0}\left(\gamma_1^L-2\gamma_2^L+1\right) \tag{7.3.7}$$

$$N=-3\frac{\hbar^2}{m_0}\gamma_3^L+\frac{P^2}{E_G} \tag{7.3.8}$$

At this point if we continued with the usual definitions for A' and P_0 we would end up with spurious solutions at large k and have states in the band gap as was the case in the two band model. Consequently we need to follow the methods in [66] to eliminate the spurious solutions. Essentially this means setting A' to zero and solving for P such that we get the correct effective mass. The main justification for this model is simply that it still reproduces the correct band structure at the gamma point. Unfortunately this only works for certain material parameters and so it is essential to check that the band structure is in fact correct when using a new material. The new fitted definition of E_P is:

$$E_P=\frac{3m_0/m_e}{2/E_G+1/(E_G+\Delta)} \tag{7.3.9}$$

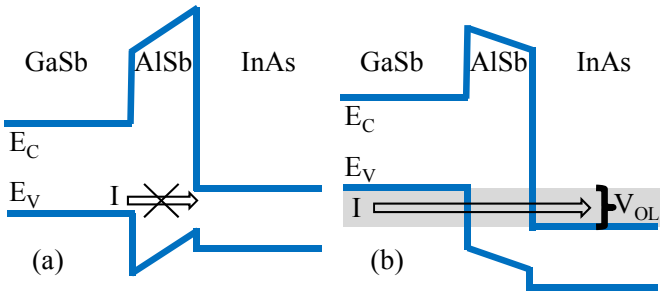


Figure 7.5: Fig 1: (a) No current can flow when the bands do not overlap. (b) Once the bands overlap, current can flow. The applied potential is dropped entirely across the AlSb barrier.

Now we have finished defining the band structure we need to convert the k- vectors into discrete spatial derivatives. At each lattice point there are four orbitals, S, X, Y, and Z to give a wavefunction of:

$$\Psi_{nx.ny} = \begin{bmatrix} S_{nx,ny} \\ X_{nx,ny} \\ Y_{nx,ny} \\ Z_{nx,ny} \end{bmatrix} \tag{7.3.10}$$

To show how to convert the k vectors to discrete differences we will give a few illustrative examples:

$$k_x X_{nx,ny} = -i\frac{d}{dx}X_{nx,ny} = \frac{-i}{2a}\left(X_{nx+1,ny} - X_{nx-1,ny}\right) \tag{7.3.11}$$

$$k_x^2 X_{nx,ny} = -\frac{d^2}{dx^2}X_{nx,ny} = \frac{-1}{a^2}\left(X_{nx+1,ny} - 2\cdot X_{nx,ny} + X_{nx-1,ny}\right) \tag{7.3.12}$$

$$k_y^2 X_{nx,ny} = -\frac{d^2}{dx^2}X_{nx,ny} = \frac{-1}{a^2}\left(X_{nx,ny+1} - 2\cdot X_{nx,ny} + X_{nx,ny-1}\right) \tag{7.3.13}$$

$$k_x k_y X_{nx,ny} = -\frac{d}{dx}\frac{d}{dy}X_{nx,ny} = \frac{-1}{4a^2}\left(X_{nx+1,ny+1} - X_{nx+1,ny-1} - X_{nx-1,ny+1} + X_{nx-1,ny-1}\right) \tag{7.3.14}$$

Using this discretization method we can now define a spatially varying Hamiltonian as done in the previous section. Since we are working in two dimensions we can leave k_z as a constant. In fact, now that the basis set is fully defined, the rest of the analysis is identical to that in the previous section.

7.4 NEGF Simulation Results

Now that NEGF Model is fully defined we can apply it to simulate 1d-$1d_{edge}$ junction. All the of the material parameters used are listed at the end of the chapter in Section 7.5

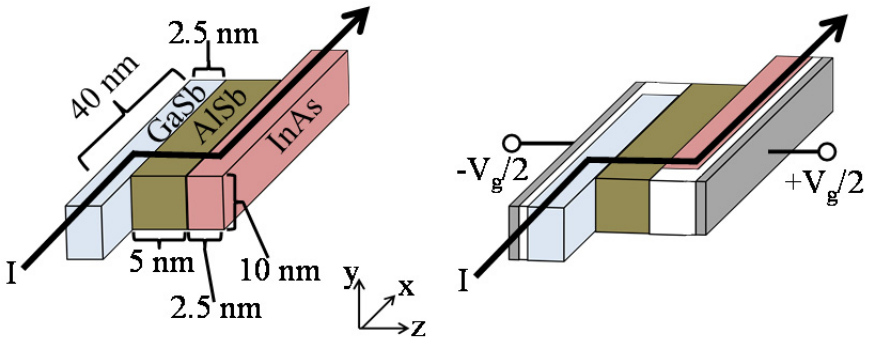


Figure 7.6: (a) We model tunneling between two coupled nanowires (1d-1d$_{edge}$) (b) Gates can be attached to each wire

One of the byproducts of pn-junction dimensionality analysis is that quantum confinement in the tunneling direction on either side of a tunnel junction greatly increases the tunneling current! This arises analytically, but here we derive added insights into the benefit of quantum confined tunneling through numerical simulation. This broadly validates the great increase in tunneling current, but reveals some new oscillatory features in the I-V curve that were previously unnoticed.

We compute the small bias conductance using 2d ballistic transport simulations within the nonequilibrium green's function (NEGF) formalism. We model the bandstructure using an 8×8 k.p Hamiltonian described in the previous section and ignore the effects of strain. We take the electronic potential to be dropped entirely across the tunneling barrier and plot the conductance as a function of the overlap potential, V_{OL}, as shown in Figure 7.5.

First we consider the tunneling between GaSb/InAs quantum wires (1d-1d$_{edge}$) pn junctions as shown in Figure 7.6. We choose the GaSb/InAs system because it has become accepted as the preferred material platform since it doesn't require heavy doping and has a favorable Type III, broken gap, band alignment. Gate electrodes can be added to the nanowires

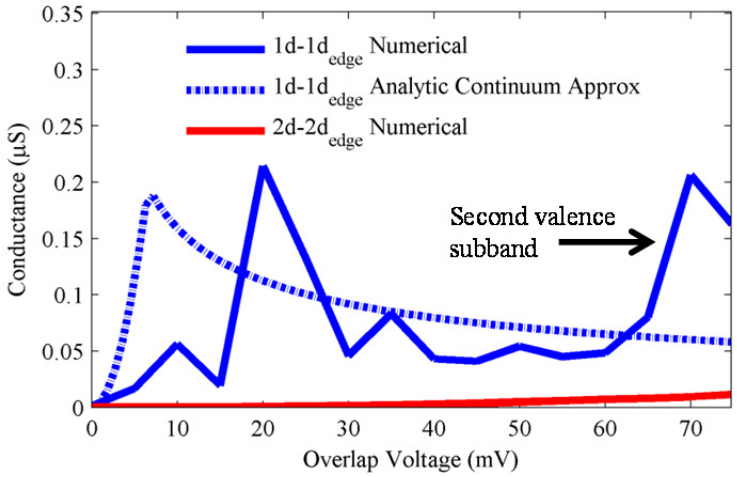


Figure 7.7: The 1d-1d$_{edge}$ conductance in is plotted as a function of V_{OL}. We see that it is 10 times larger than the 2d-2d$_{edge}$ conductance and that numerical calculation oscillates as a function of V_{OL}.

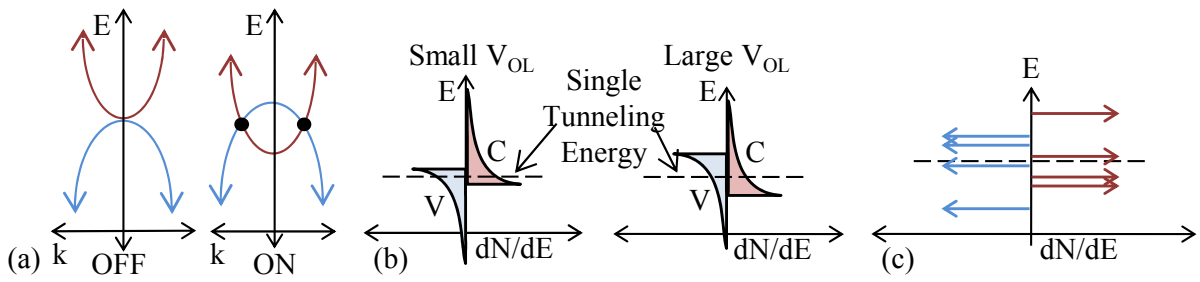


Figure 7.8: (a) There is only a single tunneling energy because of the conservation of energy and momentum. (b) The 1d density of states at the tunneling energy is different at different overlap voltages. (c) In small devices the 1d DOS is actually a series of individual states.

to control the overlap voltage V_{OL}. Assuming a continuum density of states model, the current should diverge upwards and then fall off as $1/\sqrt{V_{OL}}$ as shown by the dotted line in Figure 7.7. Conservation of energy and transverse momentum only allows current to flow at a single energy as shown in Figure 7.8a. This causes the shape of the conductivity curve to represent the 1d density of states (DOS) at the tunneling energy as shown in Figure 7.8b. However, in small devices, the 1d DOS does not form a continuum, but rather a series of individual levels as shown in Figure 7.8c. This is because the transmission probability is maximized when matching wave vectors fit in the 40 nm overlap on each side of the junction. This is illustrated in Figure 7.9. As seen in Figure 7.7, this causes the actual shape of the conductivity to be dominated by an oscillatory behavior, but only the first peak is relevant in a switching device. Except for the conductivity oscillations, the analytical continuum approximation and the more exact numerical 1d-$1d_{edge}$ curves in Figure 7.7 are similar.

Increasing the length of the overlap causes the distance between the conductance peaks to decrease as shown in Figure 7.10. In this plot we assumed ky=0. This will change the band structure but the qualitative results should be the same.

Now we introduce the 2d-$2d_{edge}$ pn junction which simply eliminates the quantum

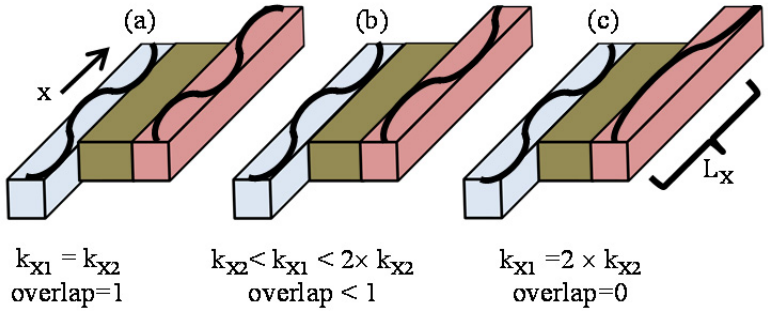


Figure 7.9: (a) When the transverse momentum, k_X, is the same in both nanowires, the transverse overlap integral is 1 and the conductivity is high (b) When k_X is slightly different on each side the transverse overlap integral starts to fall (c) When the k vectors differ by a multiple of π/L_X the overlap integral drops all the way to zero

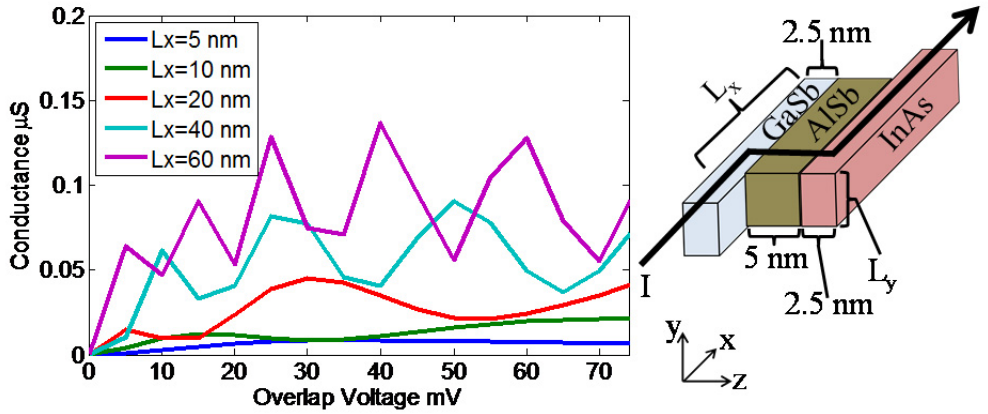


Figure 7.10: Increasing the overlap between the nanowires causes the spacing between the conductance peaks to decrease

confinement in the tunneling direction as shown in Figure 7.11. Once again gates can be added on top of the quantum wells to control the overlap potential V_{OL}. We see in Figure 7.7 that without quantum confinement in the tunneling direction (2d-2d$_{edge}$ junction) the magnitude of the conductivity is 10 times lower than the 1d-1d$_{edge}$! Counter-intuitively, shrinking the device and truncating the quantum wells increases the conductivity. This is because the quantum confinement increases the rate of tunneling attempts on both sides of the junction and improves the wavefunction overlap between each side of the junction.

By using quantum confinement along the tunneling direction, the conductivity of TFET's is significantly increased. This will help overcome the limited current drive capability of TFET's.

7.5 APPENDIX A: NEGF MODEL PARAMETERS

The material constants used are tabulated in Table 7.1. The band gaps (E_G) and valence band alignments relative to InAs (ΔE_V) are taken from [72]. The material constants from InAs are from [67]. The parameters for GaSb and AlSb are from [73]. In order to use the 8 band k.p model presented in Section 7.3, we assumed the AlSb had a direct band gap. This will not cause a significant error as the tunneling mostly occurs through the valence band due to the band

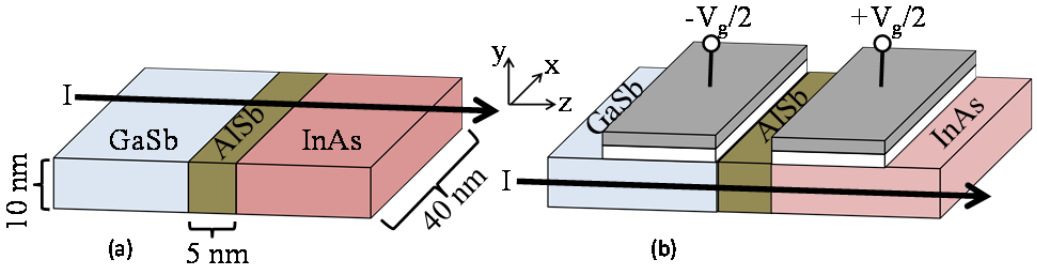


Figure 7.11: (a) We model tunneling between two quantum wells (2d-2d$_{edge}$) with the same AlSb barrier as the 1d-1d$_{edge}$ nanowires. Except for the added confinement along z, the nanowires are identical to the quantum wells. (b) A possible scheme for attaching gates to the quantum wells is shown.

alignments. The oxide is an artificial material used to block current and is based on the AlSb material parameters with a much larger band gap. Hetero-junctions are treated using average material parameters as appropriate.

	InAs	GaSb	AlSb	Oxide
E_G (eV)	0.354	0.78	1.61	12
γ_1	19.67	13.4	5.18	5.18
γ_2	8.37	4.7	1.19	1.19
γ_3	9.29	6.0	1.97	1.97
Δ (eV)	0.371	0.76	0.676	0.676
m_e/m_0	0.023	0.039	0.27	0.27
a_0 (A)	6.06	6.10	6.14	6.1355
ΔE_V (eV)	0	0.51	0.1	-6

Table 7.1: Material parameters used in the NEGF calculation

Chapter 8: Future Directions

Now that we have gone through everything that is needed to design a good tunneling switch, we present a few ideas that use this understanding to design new devices.

8.1 InAs / GaSb Quantum Well Based Structures

As the preferred tunneling geometry is tunneling between two quantum wells, we could design a structure to do exactly that. Since InAs and GaSb have very favorable band alignments, we should use that material system to design the tunneling structure. First we consider a diode structure and then we consider a full transistor structure based on the InAs/GaSb backward diode structure.

8.1.1 InAs/GaSb Quantum Well Diode

The basic structure behind a one dimensional quantum well tunneling structure is shown in Figure 8.1. Since the current is flowing in one dimension the carriers need to drop into the quantum wells. This means that the device will operate best in forward bias. In reverse bias carriers will need to be thermally excited out of the quantum wells and so the current will be suppressed. A possible implementation in the InAs/GaSb system is shown in Figure 8.2(a). Here an implementation using a GaInAsSb quaternary alloy is shown. Using the quaternary alloy allows for favorable band alignments and allows carriers to easily fall into the quantum wells. The quantum wells are undoped to eliminate the doping band tails. In order to get carriers into the wells modulation doping is used. The first two nanometers of the AlGaSb and GaInAsSb cladding layers are nominally undoped to allow for dopants to diffuse. The rest of the cladding is heavily doped. After the annealing steps some dopants will diffuse into the undoped

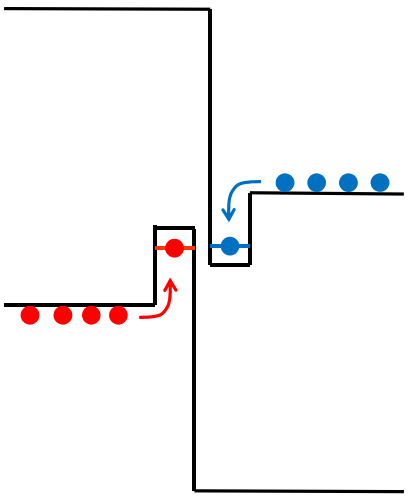


Figure 8.1: A one dimensional diode based on quantum well tunneling is shown. In forward bias carriers fall into the quantum wells and tunnel

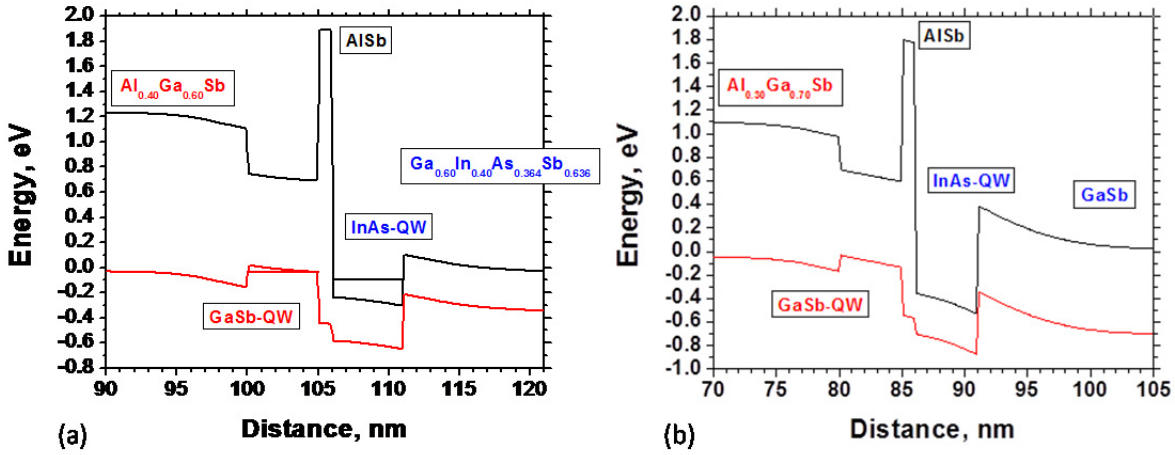


Figure 8.2: A 1d quantum well diode based on (a) quaternary (b) ternary materials is shown. The quaternary structure has more favorable band alignments, but is harder to grow.

regions, but the quantum wells will have a very low dopant concentration. Figure 8.2(b) shows a similar structure that is implemented using binary and ternary alloys. This structure is far easier to grow. However, It will be more difficult for electrons from the GaSb to enter the InAs quantum well as they need to overcome the triangular barrier between 90 and 95 nm on the figure.

Using this structure one should be able to demonstrate the 2d-2d$_{face}$ conductance increase as well as the steeper turn on.

8.1.2 InAs/GaSb Quantum Well Transistor

If we want to have reverse bias operation or if we want to design a three terminal device we need to directly contact each quantum well. This is shown in Figure 8.3(a). The current path is shown in Figure 8.3(b). The GaSb quantum well is directly contacted, current tunnels through an optional AlSb barrier into the InAs and to the contact. The Fermi level in the InAs is set by

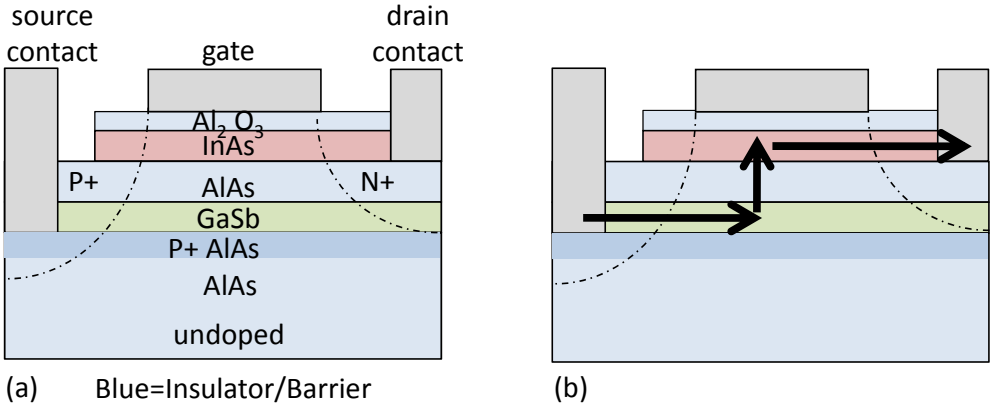


Figure 8.3: A three terminal transistor structure based on the InAs /GaSb 2d-2d$_{face}$ structure is shown. (a) The layer structure is shown (b) The current path is shown.

Figure 8.4: (from Tsai 1994) The band diagram of a resonant interband tunneling diode is shown

the gate work function and bias. The GaSb Fermi level is set by modulation doping in the lower AlSb barrier layer. Direct conduction through the GaSb layer is suppressed by forming a reverse biased PN junction between the source and drain contacts. Direct conduction in the InAs is suppressed by etching away the InAs prior to forming the source contact. This type of a quantum well structure should result in a high performance TFET.

8.2 RESONANT INTERBAND TUNNELING BACKWARD DIODES

So far, our quantum well structures have relied on either directly contacting the quantum wells or allowing carriers to scatter into the quantum wells. However, it is possible to also have a tunneling contact into the quantum wells. This would result is would result in a resonant interband tunneling diode (RITD) [65, 74, 75]. These devices have given record peak to valley current ratios as high as 144 [74]in Esaki diodes. The band diagram of an RITD is shown in Figure 8.4.

Since the tunneling occurs between two quantum wells the same conductance boost predicted for a 2d-2d_{face} junction should occur. It is also possible that the resonant tunneling provides an added benefit and should be investigated further.

8.3 2D-2D WORK-FUNCTION CONTROLLED TUNNELING FIELD EFFECT TRANSISTOR

It is also possible to take advantage of the benefits of quantum well tunneling by using a homojunction and gate workfunctions to get the desired band alignment. A basic 2d-2d work-function controlled TFET is illustrated in Figure 8.5. Figure 8.5 shows a FinFET structure and Figure 8.6 shows a layered structure. The direction of current flow is given in Figure 8.7. There are many different ways to physically implement this type of structure [76-84]. It is possible to create a variety of layered structures [76-78] or it can be made as a variation of a FinFET or tri-gate structure [79-84]. To create a TFET the source needs to be doped p-type and the drain

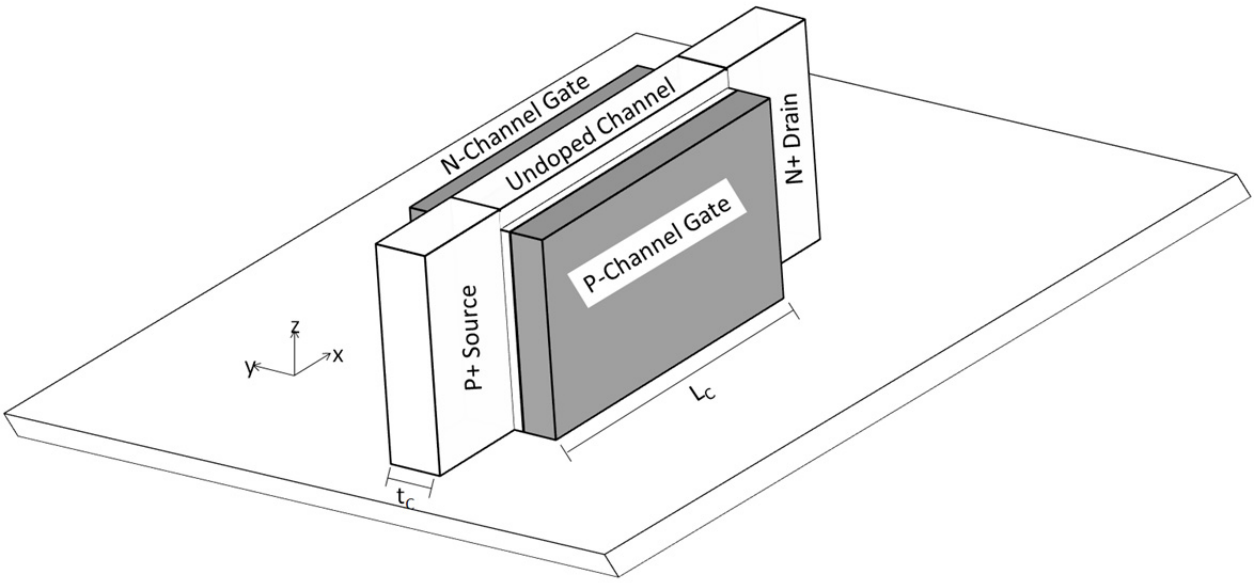


Figure 8.5: A double gate dual channel TFET in a FinFET configuration is shown. The gates should have different work functions to induce both an n and a p channel.

needs to be doped n type. The N-Channel Gate is composed of a metal and high-k dielectric with a low work function such that the surface of the channel becomes n-type. The P-Channel Gate has a large work function such that the surface of the channel near the P Channel Gate becomes p-type. To create a complementary device for CMOS logic the control gate simply needs to be reversed. Figure 8.8 shows the resulting band diagram across the channel along the y-direction. The channel should be undoped to allow gate work function control to form an n-channel and a p-channel on opposite faces. This is very different from previous double gate proposals[85], as both the n and p channels are needed simultaneously. Leaving the channel undoped also preserves the material quality which results in sharper band edges and allows the

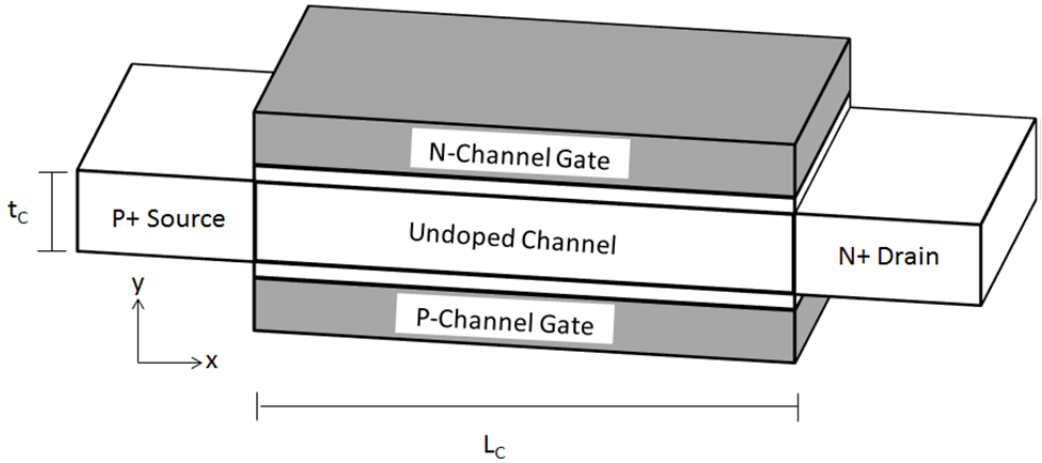


Figure 8.6: A double gate dual channel TFET is shown. It is turned on its side to make a layered structure.

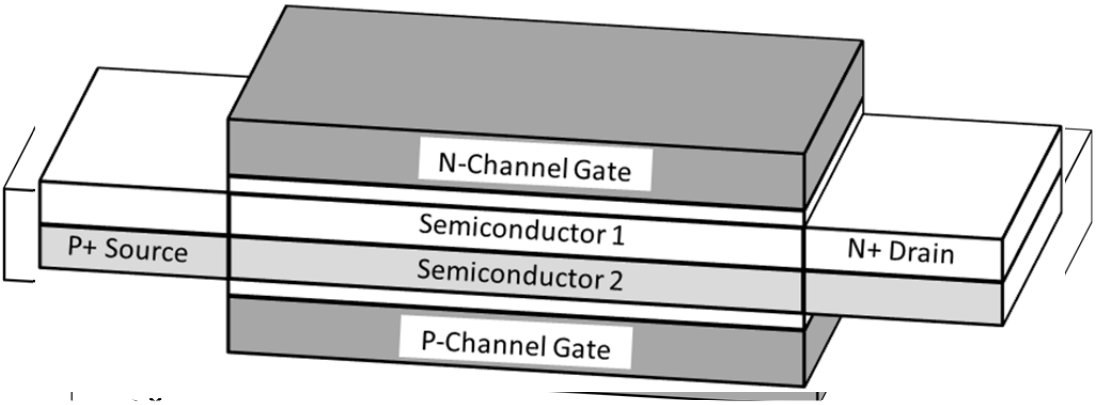


Figure 8.9: The channel can incorporate a semiconductor heterojunction.

Figure 8.7: The direction of current flow is shown. The current flows along the -x direction but tunnels along the -y-direction.

potential for the device to have a sharper, lower voltage, turn-on. Tunneling will occur over the broad area of overlap between the two gates resulting in an increased face-to-face conductivity.

Any semiconductor can be used for the channel, source and drain. Possibilities include silicon, germanium, GaAs, InAs, GaSb and so on. Using silicon or germanium would make the structure compatible with current CMOS processes. The FinFET geometry is particularly attractive, in that the main change would be opposite "gate stacks" on either side of the Fin, versus identical stacks on either side as we have for conventional FinFET's. The channel thickness, $t_{C,}$ should be thin enough to allow tunneling from face-to-face. The channel length, L_C, can be varied based on the desired device area. The gate oxide must be part of a "gate stack" that includes Work Function control, so that one face is n-type and the other face is p-type.

It's also possible to use a heterojunction in the channel to increase the tunneling

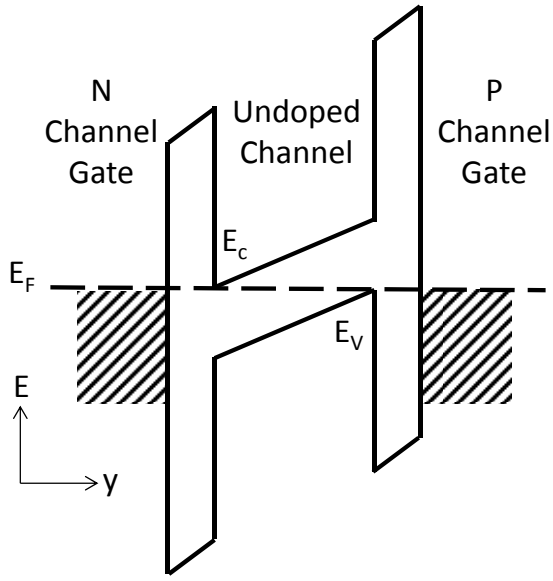


Figure 8.8: The band diagram across the channel is shown. Work Function control tilts the bands as shown. Tunneling occurs across the channel along the y direction.

probability and decrease the required workfunction difference. This is shown in Figure 8.9. For instance, semiconductor 1 could be silicon and semiconductor 2 could be germanium.

8.3.1 Using Source/Drain Extensions to Suppress Unwanted Tunneling

In order to suppress unwanted tunneling and have a sharper low voltage turn-on, it is possible to use source/drain extensions as shown in Figure 8.10. The two optional extension regions (Ext.) shown in Figure 8.10 should be undoped or lightly doped. The source extension can be lightly p doped and the drain extension can be lightly n doped, and can be produced as part of a self-aligned process, if desired. These extensions eliminate any direct tunneling to the source or drain and ensure that the tunneling occurs in the channel region.

8.3.2 Using Semiconductor Contact for the P Channel Gate

An alternative to the gate stack is the use of a semiconductor heterostructure on one or both sides of the undoped semiconductor film as shown in Figure 8.11. The band diagram across the channel along the y-direction is shown in Figure 8.12. The basic idea is to replace the P-Channel Gate with a HEMT (High Electron Mobility Transistor) style gate contact. The cladding layer should be a different semiconductor than the channel with large valence band offset as shown in Figure 8.12. The optional source drain extensions can be formed by using standard processing techniques such as depositing spacers prior to implanting the source and drain. Any dielectric such as silicon nitride can be used for the spacers.

One possible material system is to use germanium for the channel/source/drain and silicon for the cladding layer. The large band offsets between silicon and germanium prevent tunneling between the cladding layer and the drain and ensures that the current conduction occurs entirely within the germanium. To further suppress unwanted tunneling the portion of the cladding layer under the drain can be left undoped or lightly doped.

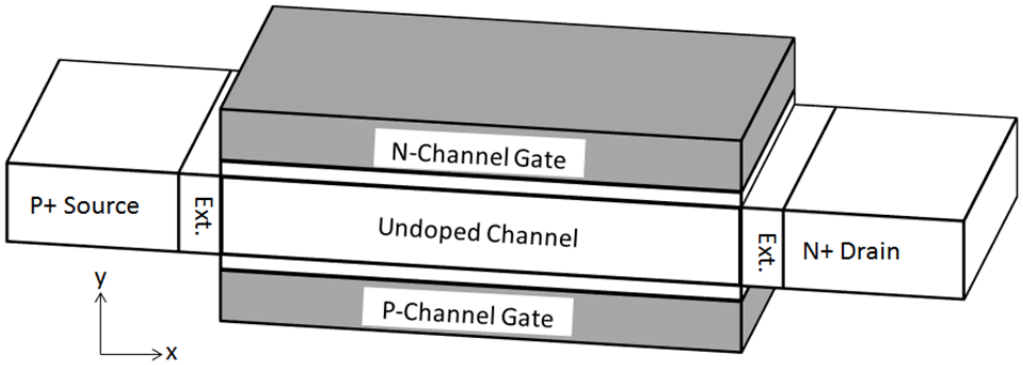


Figure 8.10: Unwanted tunneling can be suppressed by including lightly doped source and drain extensions

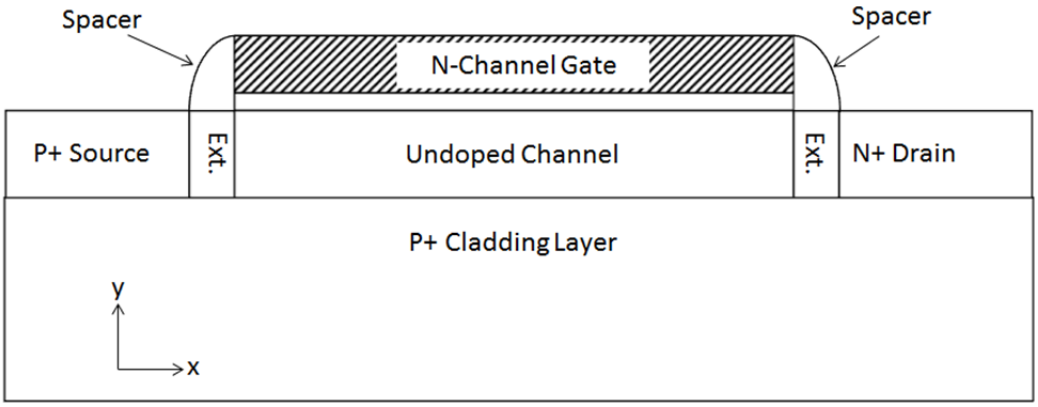


Figure 8.11: A P+ cladding layer can be used in place of a second gate to induce a p+ channel

If the film structure has a "gate-stack" on one side, and a semiconductor heterostructure on the other side, the complementary device will require a different set of materials as shown in Figure 8.13. In this case a gate metal with a high work function should be used. The channel could be silicon and the cladding layer could be germanium or a silicon-germanium alloy

8.3.3 Conclusion

These structures overcome the limitations of previous TFET designs by taking advantage of a large tunneling area, the unique physics of 2d-2d$_{face}$ tunneling and an undoped junction to provide a high on current and potentially steep turn-on. Depending on the materials chosen, these designs are compatible with current CMOS process technology. This will allow for the easy adoption by industry and result in a reduction in the power consumption of current electronics.

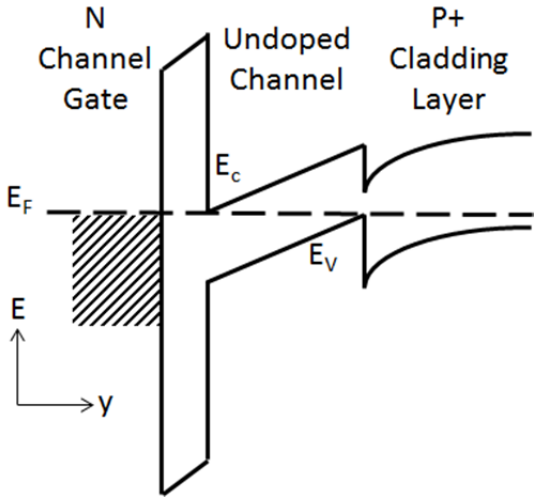


Figure 8.12: The band diagram including a cladding layer is shown. The cladding layer has a large valence band offset and uses modulation doping to induce a p-type channel near the cladding layer.

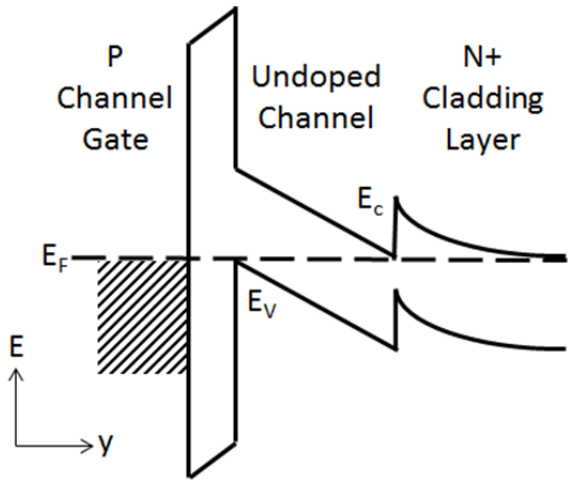


Figure 8.13: The band diagram of the complementary device is shown. To form a complementary device the cladding layer should have a large conduction band offset and be doped n-type. The gate should also have a large work function to induce a p-type channel near the gate

8.4 CONCLUSION

As we have seen, there are a variety of possible devices that need to be investigated further and that may enable the tunneling switch to fulfill its promise. Throughout this dissertation we have seen all of the various effects that can both limit and enhance the performance of a tunneling switch. By combining these ideas and creating a device such as those suggested in this section, TFETs may be able to revolutionize electronics.

References

[1] J. G. Koomey, "Growth in data center electricity use 2005 to 2010," Oakland, CA 2011.

[2] (2011). *The International Technology Roadmap for Semiconductors (ITRS)*. Available: http://www.itrs.net/

[3] S. H. Kim, *et al.*, "Germanium-source tunnel field effect transistors with record high I/sub ON//I/sub OFF," presented at the 2009 Symposium on VLSI Technology, Kyoto, Japan, 2009.

[4] W. Y. Choi, *et al.*, "Tunneling Field-Effect Transistors (TFETs) With Subthreshold Swing (SS) Less Than 60 mV/dec," *IEEE Electron Device Letters,* vol. 28, pp. 743-745, Aug 2007.

[5] K. Jeon, *et al.*, "Si tunnel transistors with a novel silicided source and 46mV/dec swing," presented at the 2010 IEEE Symposium on VLSI Technology, Honolulu, Hawaii, 2010.

[6] T. Krishnamohan, *et al.*, "Double-gate strained-Ge heterostructure tunneling FET (TFET) With record high drive currents and < 60 mV/dec subthreshold slope," presented at the IEDM 2008. IEEE International Electron Devices Meeting, San Francisco, CA,, 2008.

[7] J. Knoch, *et al.*, "Impact of the dimensionality on the performance of tunneling FETs: Bulk versus one-dimensional devices," *Solid-State Electronics,* vol. 51, pp. 572-578, Apr 2007.

[8] A. C. Seabaugh and Q. Zhang, "Low-Voltage Tunnel Transistors for Beyond CMOS Logic," *Proceedings of the Ieee,* vol. 98, pp. 2095-2110, Dec 2010.

[9] J. Appenzeller, *et al.*, "Band-to-Band Tunneling in Carbon Nanotube Field-Effect Transistors," *Physical Review Letters,* vol. 93, p. 196805, 2004.

[10] R. Gandhi, *et al.*, "CMOS-Compatible Vertical-Silicon-Nanowire Gate-All-Around p-Type Tunneling FETs With <= 50-mV/decade Subthreshold Swing," *Ieee Electron Device Letters,* vol. 32, pp. 1504-1506, Nov 2011.

[11] G. Dewey, *et al.*, "Fabrication, characterization, and physics of III-V heterojunction tunneling Field Effect Transistors (H-TFET) for steep sub-threshold swing," *2011 IEEE International Electron Devices Meeting (IEDM 2011),* pp. 33.6 (4 pp.)-33.6 (4 pp.)33.6 (4 pp.), 2011 2011.

[12] J. N. Schulman and D. H. Chow, "Sb-heterostructure interband backward diodes," *IEEE Electron Device Letters,* vol. 21, pp. 353-355, Jul 2000.

[13] S. M. Sze and K. K. Ng, *Physics of semiconductor devices*: Wiley-Interscience, 2007.

[14] R. G. Meyers, *et al.*, "Bias and temperature dependence of Sb-based heterostructure millimeter-wave detectors with improved sensitivity," *Ieee Electron Device Letters,* vol. 25, pp. 4-6, Jan 2004.

[15] J. Karlovsky and A. Marek, "On an Esaki Diode Having Curvature Coeficient Greater Than E/KT," *Czechoslovak Journal of Physics,* vol. 11, pp. 76-&, 1961.

[16] J. Karlovsky, "Curvature Coefficient of Germanium Tunnel and Backward Diodes," *Solid-State Electronics,* vol. 10, pp. 1109-1111, 1967.

[17] N. Su, *et al.*, "Sb-Heterostructure Millimeter-Wave Detectors With Reduced Capacitance and Noise Equivalent Power," *Electron Device Letters, IEEE,* vol. 29, pp. 536-539, 2008.

[18] M. J. Tadjer, *et al.*, "On the high curvature coefficient rectifying behavior of nanocrystalline diamond heterojunctions to 4H-SiC," *Applied Physics Letters,* vol. 97, Nov 2010.
[19] Z. Zhang, *et al.*, "Sub-micron Area Heterojunction Backward Diode Millimeter-wave Detectors With 0.18 pW/Hz 1/2 Noise Equivalent Power," *IEEE Microwave and Wireless Components Letters,* vol. 21, pp. 267-269, May 2011.
[20] N. Su, *et al.*, "Temperature dependence of high frequency and noise performance of Sb-heterostructure millimeter-wave detectors," *IEEE Electron Device Letters,* vol. 28, pp. 336-339, May 2007.
[21] T. Y. Chan, *et al.*, "The Impact of Gate-Induced Drain Leakage Current on MOSFET Scaling," in *1987 International Electron Devices Meeting*, 1987, pp. 718-721.
[22] E. O. Kane, "Theory of Tunneling," *Journal of Applied Physics,* vol. 32, pp. 83-91, 1961.
[23] N. Holonyak, *et al.*, "Direct Observation of Phonons During Tunneling in Narrow Junction Diodes," *Physical Review Letters,* vol. 3, pp. 167-168, 1959.
[24] S. L. Chuang, *Physics of Optoelectronic Devices*. New York: John Wiley & Sons, Inc, 1995.
[25] G. D. Cody, "Urbach Edge of Crystalline and Amorphous Silicon - A Personal Review," *Journal of Non-Crystalline Solids,* vol. 141, pp. 3-15, Mar 1992.
[26] T. Tiedje, *et al.*, "Limiting Efficiency of Silicon Solar Cells," *IEEE Transactions on Electron Devices,* vol. 31, pp. 711-716, 1984.
[27] S. O. Koswatta, *et al.*, "Band-to-Band Tunneling in a Carbon Nanotube Metal-Oxide-Semiconductor Field-Effect Transistor Is Dominated by Phonon-Assisted Tunneling," *Nano Letters,* vol. 7, pp. 1160-1164, 2007/05/01 2007.
[28] S. O. Koswatta, *et al.*, "Influence of phonon scattering on the performance of p-i-n band-to-band tunneling transistors," *Applied Physics Letters,* vol. 92, Jan 2008.
[29] S. O. Koswatta, *et al.*, "Nonequilibrium green's function treatment of phonon scattering in carbon-nanotube transistors," *Ieee Transactions on Electron Devices,* vol. 54, pp. 2339-2351, Sep 2007.
[30] M. Luisier and G. Klimeck, "Simulation of nanowire tunneling transistors: From the Wentzel-Kramers-Brillouin approximation to full-band phonon-assisted tunneling," *Journal of Applied Physics,* vol. 107, Apr 2010.
[31] Y. Yoon, *et al.*, "Role of phonon scattering in graphene nanoribbon transistors: Nonequilibrium Green's function method with real space approach," *Applied Physics Letters,* vol. 98, May 2011.
[32] M. Luisier and G. Klimeck, "Atomistic full-band simulations of silicon nanowire transistors: Effects of electron-phonon scattering," *Physical Review B,* vol. 80, Oct 2009.
[33] S. Adachi, *Handbook on Physical Properties of Semiconductors.* vol. 1: Springer - Verlag, 2004.
[34] R. P. Feynman, *et al.*, *The Feynman Lectures on Physics* vol. II. Reading, Massachusetts: Addison Wesley Publishing Company, 1964.
[35] C. Kittel, *Introduction to Solid State Physics*, 8 ed.: John Wiley & Sons, Inc., 2004.
[36] C. G. Van De Walle, "Band Lineups and Deformation Potentials in the Model-Solid Theory," *Physical Review B,* vol. 39, pp. 1871-1883, Jan 1989.
[37] C. Herring and E. Vogt, "Transport and Deformation-Potential Theory for Many-Valley Semiconductors with Anisotropic Scattering," *Physical Review,* vol. 101, pp. 944-961, 1956.

[38] I. Balslev, "Influence of Uniaxial Stress on the Indirect Absorption Edge in Silicon and Germanium," *Physical Review,* vol. 143, p. 636, 1966.
[39] R. Vrijen*, et al.*, "Electron-spin-resonance transistors for quantum computing in silicon-germanium heterostructures," *Physical Review A,* vol. 62, p. 012306, 2000.
[40] C. Penn*, et al.*, "Energy Gaps and Band Structure of SiGe and their Temperature Dependence," in *Properties of Silicon Germanium and SiGe:Carbon*, E. Kasper and K. Lyutovich, Eds., ed London: The Institution of Electrical Engineers, 2000.
[41] C. G. Van de Walle and R. M. Martin, "Theoretical calculations of heterojunction discontinuities in the Si/Ge system," *Physical Review B,* vol. 34, p. 5621, 1986.
[42] E. Daub and P. Wurfel, "Ultra-Low Values of the Absorption Coefficient for Band--Band Transitions in Moderately Doped Si Obtained From Luminescence," *Journal of Applied Physics,* vol. 80, pp. 5325-5331, 1996.
[43] E. O. Kane, "Thomas-Fermi Approach to Impure Semiconductor Band Structure," *Physical Review,* vol. 131, pp. 79-88, 1963.
[44] E. O. Kane, "Band tails in semiconductors," *Solid-State Electronics,* vol. 28, pp. 3-10, 1985.
[45] B. I. Halperin and M. Lax, "Impurity-Band Tails in the High-Density Limit. I. Minimum Counting Methods," *Physical Review,* vol. 148, pp. 722-740, 1966.
[46] B. I. Halperin and M. Lax, "Impurity-Band Tails in the High-Density Limit. II. Higher Order Corrections," *Physical Review,* vol. 153, pp. 802-814, 1967.
[47] V. Sa-yakanit*, et al.*, "Impurity-band density of states in heavily doped semiconductors: Numerical results," *Physical Review B,* vol. 25, pp. 2776-2780, 1982.
[48] W. B. Jackson*, et al.*, "Density of gap states of silicon grain boundaries determined by optical absorption," *Applied Physics Letters,* vol. 43, pp. 195-197, 1983.
[49] P. M. Solomon*, et al.*, "Universal tunneling behavior in technologically relevant P/N junction diodes," *Journal of Applied Physics,* vol. 95, pp. 5800-5812, 2004.
[50] S. Mookerjea*, et al.*, "Temperature-Dependent I-V Characteristics of a Vertical In(0.53)Ga(0.47)As Tunnel FET," *Ieee Electron Device Letters,* vol. 31, pp. 564-566, Jun 2010.
[51] R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, 3rd ed.: John Wiley and Sons, 2003.
[52] G. A. M. Hurkx*, et al.*, "A new recombination model for device simulation including tunneling," *Electron Devices, IEEE Transactions on,* vol. 39, pp. 331-338, 1992.
[53] J. Furlan, "Tunnelling generation-recombination currents in a-Si junctions," *Progress in Quantum Electronics,* vol. 25, pp. 55-96, 2001.
[54] G. A. M. Hurkx*, et al.*, "A new recombination model describing heavy-doping effects and low-temperature behaviour," in *Electron Devices Meeting, 1989. IEDM '89. Technical Digest., International*, 1989, pp. 307-310.
[55] K. Ganapathi*, et al.*, "Analysis of InAs vertical and lateral band-to-band tunneling transistors: Leveraging vertical tunneling for improved performance," *Applied Physics Letters,* vol. 97, pp. 033504-3, 2010.
[56] P. Patel, "Steep Turn On/Off "Green" Tunnel Transistors," PhD, Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, 2010.
[57] J. Bardeen, "Tunneling From a Many Particle Point of View," *Physical Review Letters,* vol. 6, pp. 57-59, 1961.
[58] C. B. Duke, *Tunneling in Solids*. New York: Academic Press, Inc, 1969.

[59] W. A. Harrison, "Tunneling from an independent-particle point of view," *Physical Review,* vol. 123, pp. 85-89, 1 July 1961.
[60] J. R. Oppenheimer, "Three notes on the quantum theory of aperiodic effects," *Physical Review,* vol. 31, pp. 66-81, Jan 1928.
[61] G. Gamow, "Zur Quantentheorie des Atomkernes," *Z. Physik,* vol. 51, p. 204, 1928.
[62] M. A. Kastner, "The single-electron transistor," *Reviews of Modern Physics,* vol. 64, p. 849, 1992.
[63] S. Datta, *Quantum Transport : Atom to Transistor*. Cambridge, UK: Cambridge University Press, 2005.
[64] D. J. Griffiths, *Introduction to Quantum Mechanics*, 2 ed. Upper Saddle River, NJ: Prentice Hall, Inc., 1994.
[65] M. Sweeny and J. M. Xu, "Resonant Interband Tunnel-Diodes," *Applied Physics Letters,* vol. 54, pp. 546-548, Feb 1989.
[66] B. A. Foreman, "Elimination of spurious solutions from eight-band k.p theory," *Physical Review B,* vol. 56, pp. R12748-R12751, 1997.
[67] D. Gershoni*, et al.*, "Calculating the optical properties of multidimensional heterostructures: Application to the modeling of quaternary quantum well lasers," *Quantum Electronics, IEEE Journal of,* vol. 29, pp. 2433-2450, 1993.
[68] M. P. L. Sancho*, et al.*, "Highly convergent schemes for the calculation of bulk and surface Green functions," *Journal of Physics F: Metal Physics,* vol. 15, p. 851, 1985.
[69] E. O. Kane, "Zener Tunneling in Semiconductors," *Journal of Physics and Chemistry of Solids,* vol. 12, pp. 181-188, 1959.
[70] T. B. Bahder, "Eight-band k.p model of strained zinc-blende crystals," *Physical Review B,* vol. 41, pp. 11992-12001, 1990.
[71] P. Enders*, et al.*, "k.p theory of energy bands, wave functions, and optical selection rules in strained tetrahedral semiconductors," *Physical Review B,* vol. 51, pp. 16695-16704, 1995.
[72] H. Kroemer, "The 6.1 Å family (InAs, GaSb, AlSb) and its heterostructures: a selective review," *Physica E: Low-dimensional Systems and Nanostructures,* vol. 20, pp. 196-203, 2004.
[73] I. Vurgaftman*, et al.*, "Band parameters for III-V compound semiconductors and their alloys," *Journal of Applied Physics,* vol. 89, pp. 5815-5875, Jun 2001.
[74] H. H. Tsai*, et al.*, "P-N double quantum well resonant interband tunneling diode with peak-to-valley current ratio of 144 at room temperature," *Electron Device Letters, IEEE,* vol. 15, pp. 357-359, 1994.
[75] J. H. Smet*, et al.*, "Peak-to-valley current ratios as high as 50:1 at room temperature in pseudomorphic In[sub 0.53]Ga[sub 0.47]As/AlAs/InAs resonant tunneling diodes," *Journal of Applied Physics,* vol. 71, pp. 2475-2477, 1992.
[76] T. Tanaka*, et al.*, "Ultrafast Operation of V_{th}-Adjusted P^+ - N^+ Double Gate SOI MOSFETs," *IEEE Electron Device Letters,* vol. 15, pp. 386-388, Oct 1994.
[77] T. Tanaka*, et al.*, "Analysis of p + poly Si double-gate thin-film SOI MOSFETs," *International Electron Devices Meeting 1991. Technical Digest (Cat. No.91CH3075-9),* pp. 683-686686, 1991 1991.
[78] K. W. Guarini*, et al.*, "Triple-self-aligned, planar double-gate MOSFETs: devices and circuits," in *Electron Devices Meeting, 2001. IEDM Technical Digest. International*, 2001, pp. 19.2.1-19.2.4.

[79] H. Takato*, et al.*, "Impact of surrounding gate transistor (SGT) for ultra-high-density LSI's," *Electron Devices, IEEE Transactions on,* vol. 38, pp. 573-578, 1991.
[80] B. Goebel*, et al.*, "Fully depleted surrounding gate transistor (SGT) for 70 nm DRAM and beyond," in *Electron Devices Meeting, 2002. IEDM '02. International*, 2002, pp. 275-278.
[81] J. M. Hergenrother*, et al.*, "The Vertical Replacement-Gate (VRG) MOSFET: a 50-nm vertical MOSFET with lithography-independent gate length," in *Electron Devices Meeting, 1999. IEDM Technical Digest. International*, 1999, pp. 75-78.
[82] H. Gossner*, et al.*, "Vertical MOS technology with sub-0.1 μm channel lengths," *Electronics Letters,* vol. 31, pp. 1394-1396, 1995.
[83] D. Hisamoto*, et al.*, "FinFET-a self-aligned double-gate MOSFET scalable to 20 nm," *Electron Devices, IEEE Transactions on,* vol. 47, pp. 2320-2325, 2000.
[84] B. S. Doyle*, et al.*, "High Performance Fully-Depleted Tri-Gate CMOS Transistors," *IEEE Electron Device Letters,* vol. 24, pp. 263-265, Apr 2003.
[85] K. Boucart and A. M. Ionescu, "Double-Gate Tunnel FET With High-K Gate Dielectric," *Electron Devices, IEEE Transactions on,* vol. 54, pp. 1725-1733, 2007.