Designing a Low Voltage, High Current Tunneling Transistor

Sapan Agarwal
Eli Yablonovitch

Electrical Engineering and Computer Sciences
University of California at Berkeley

December 31, 2013

Acknowledgement

Designing a Low Voltage, High Current Tunneling Transistor

Sapan Agarwal, Eli Yablonovitch

4.1   Introduction

Tunneling Field Effect Transistors (TFETs) have the potential to achieve a low operating voltage by overcoming the thermally limited subthreshold swing voltage of 60mV/decade[1], but results to date have been unsatisfying. The low voltage operation is parameterized by the voltage required to obtain a 10× change in output current, called the subthreshold swing voltage, SS. The best reported subthreshold swing voltage has been measured at a low current density of ~1nA/μm, but unfortunately becomes significantly larger as the current increases. When trying to design a new low voltage switch to replace the transistor, there are three major requirements need to be fulfilled:

- The subthreshold swing voltage needs to be much steeper than 60 mV/decade and ideally only a few milli-volts per decade to reduce the operating voltage.
- A large On/Off ratio of around $10^6/1$ is needed to suppress leakage currents
- A high conductance density around 1mS/μm (or 1mA/μm at 1Volt) is needed so that the switch can be significantly smaller than the wire that it drives while maintaining a high speed.

While devices have been built that meet one or two of the three requirements, to date, no logic switch meets all three requirements [1, 2]. No one has achieved a steep subthreshold swing voltage at a high conductance.

To understand this, we first consider a simple tunneling diode in Sections 4.2-4.6 to understand the essential physics of tunneling and then in Sections 4.7-4.9 we consider the additional complexities of building a full transistor. In TFET's the challenge is complicated by the existence of two switching mechanisms. The gate voltage can be used to modulate the tunneling barrier thickness and thus the tunneling probability [3-6] as shown Fig. 4.1(a-b). The thickness of the tunneling barrier can be controlled by changing the electric field in the tunneling junction. Alternatively, it is also possible use energy filtering or density of states switching as illustrated in Fig. 4.1(c-d). If the conduction and valence band don't overlap, no current can flow. Once they do overlap, current can flow.

In Section 4.2 we analyze the tunneling barrier thickness modulation mechanism. In Section 4.3 we analyze the energy filtering switch or density of states switch. After presenting the two switching mechanisms, we analyze in Section 4.4 the existing device data which shows that experimental performance is still far worse than the Boltzmann limit 60mV/decade. This is not even close to achieving a steep subthreshold swing voltage at the current densities of interest. We propose some solutions to achieve better steepness in Section 4.5. To fulfill all three switch requirements, we introduce in Section 4.6 the benefits of quantum confinement or dimensionality. Up until this point in the chapter, the switch has been analyzed with respect to its two-terminal properties. In Section 4.7, we consider the voltage, subthreshold swing and conductance of a full TFET. In Section 4.8 we analyze the relatively poor gate efficiency which leads to additional unfortunate tradeoffs and in Section 4.9 we consider what additional effects can limit the TFET performance.
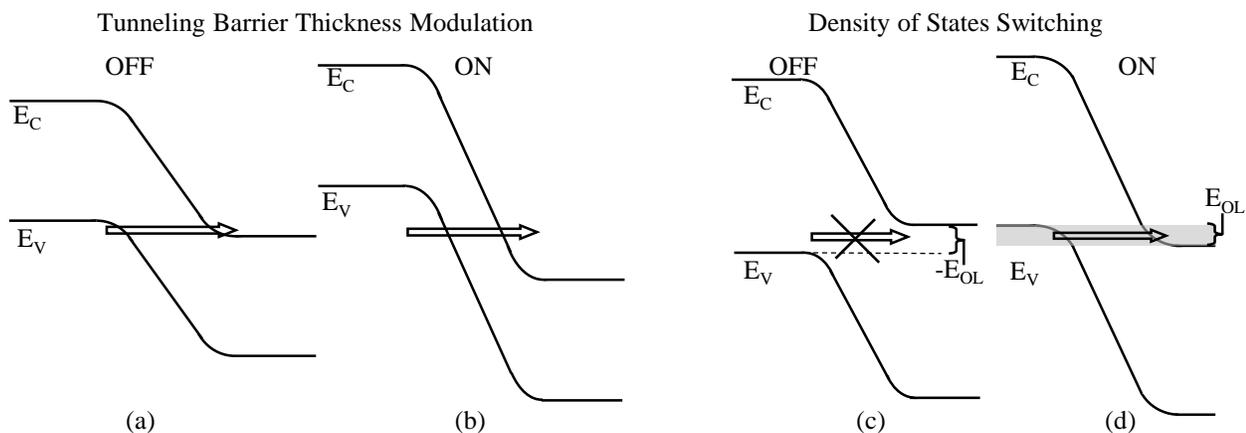
Fig. 4.1: The two different methods for achieving a steep tunneling transition are illustrated. First (a, b), the thickness of the tunneling barrier can be changed by changing the gate voltage and thus the electric field across the tunneling junction. Second (c, d), the alignment of the conduction and valence band can be used to cut off the available states for current to flow into.

4.2 Tunneling barrier thickness modulation steepness

First we consider the tunneling barrier thickness modulation mechanism. Applying a voltage bias across a tunneling junction can modulate the tunneling barrier thickness and thus the tunneling probability [3-6]. This is illustrated in Fig. 4.1(a-b). The thickness of the tunneling barrier can be controlled by changing 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 voltage bias cannot control the barrier width effectively. This results in a poor subthreshold swing voltage at high conductivities. Consequently, we will now show that the tunnel barrier thickness modulation mechanism is incapable of achieving a steep subthreshold swing voltage at the required high current density:

To estimate how steep of a turn on this can give, we need to determine how many millivolts change in potential across the barrier, ϕ, it takes to change the tunneling probability, $\top$, by a decade. Consequently, we define:

$$\frac{1}{S_{tunnel}} \equiv \frac{d \log(\top)}{d\phi} \tag{4.2.1}$$

S_{tunnel} is the tunnel swing voltage in mV/decade resulting from tunneling barrier thickness modulation. The tunneling probability, $\top$, is[7]:

$$\top(\vec{F}) = \exp\left(\frac{-\pi(m^*_{tunnel})^{1/2} E_G^{3/2}}{2\sqrt{2}\hbar q F}\right) \equiv \exp\left(\frac{-\alpha}{F}\right) \tag{4.2.2}$$

For simplicity, we will assume that the electric field across the tunneling junction, F, is constant and equal to the peak electric field. The effective mass for tunneling [1, 8, 9] is m^*_{tunnel} [1], and E_G is the band gap. All of the parameters can be collected into a single constant, α. Regardless of the exact shape of the barrier, there will be a constant α such that $\top = \exp(-\alpha/F)$. Combining Eq. (4.2.1) and Eq. (4.2.2) gives:

$$\frac{1}{S_{tunnel}} = \left(\log(e) \times \alpha \times \frac{1}{|\vec{F}|^2} \times \frac{d\vec{F}(\phi)}{d\phi}\right) = \left|\log(\top) \times \frac{dF(\phi)/d\phi}{F}\right| \tag{4.2.3}$$

To simplify we solved Eq. (4.2.2) for F in terms of $log(\top)$.

Next we need to evaluate $F/(dF(\phi)/d\phi)$. For a doped pn-junction the potential will be parabolic and so:

$$\frac{F}{dF(\phi)/d\phi} = 2\phi \tag{4.2.4}$$

In a MOSFET channel the voltage typically decays exponentially and is set by a screening length. This results in:

$$\frac{F}{dF(\phi)/d\phi} = \phi \tag{4.2.5}$$

In transistor structures such as a bilayer TFET[10, 11], the electric field is a constant and determined by the bias across the gates. Consequently:

---

[1] The tunneling mass can be computed from [8]:

$$m^*_{tunnel} = 2\left(\frac{1}{m^*_{e,z}} + \frac{1}{m^*_{h,z}}\right)^{-1}$$

The WKB model and reduced mass work well in InAs where there are carriers in the conduction band tunneling to a single valence band. However, in silicon and germanium the band gap is indirect and there are many interacting bands and so the WKB model breaks down [9]. Consequently, we use an experimentally fitted tunneling effective mass derived in [1]. While in [1] a single band tunneling model was used, we used a two band tunneling model and consequently we need to adjust the mass accordingly:

$$m^*_{2\,Band} = \left(\frac{2\sqrt{2}}{\pi} \times \frac{4\sqrt{2}}{3}\right)^2 m^*_{1\,Band}$$

This gives m*$_{tunnel}$=0.043 in InAs and 0.46 in Si.

$$\frac{F}{dF(\phi)/d\phi} = \phi \qquad\qquad (4.2.6)$$

In the best case, $F/(dF(\phi)/d\phi) = \phi$. This gives:

$$S_{tunnel} \approx \left| \frac{\phi}{\log(\mathsf{T})} \right| \qquad\qquad (4.2.7)$$

As we can see from Eq. (4.2.7), the lower the tunneling probability, the steeper the subthreshold swing voltage. This simple equation is likely to be the explanation of all the experimental steep subthreshold swing voltages that have been measured at extremely low current densities to date [3, 5, 12]! Since the steepness gets worse at high tunnel probability or higher currents, a steep subthreshold swing voltage at low currents is insufficient for making a practical logic switch. For a reasonable on-state conductance, the tunneling probability should typically be >1%.

To align the conduction and valence bands, the on-state φ must be equal to at least the band gap of the semiconductor. Consequently, the tunnel swing voltage S_{tunnel} is too large for different semiconductors: For T=1% in silicon, S_{tunnel}=560 mV/decade at the on-state. In InAs, $S_{tunnel} = 177$ mV/decade. These are worse than Boltzmann. Clearly, controlling the barrier thickness of a homojunction will not give a subthreshold swing voltage steeper than 60 mV/decade at high current densities.

To get a subthreshold swing voltage steeper than 60 mV/decade while maintaining a high on-state current, we need to reduce φ_s to less than 120 mV. This means that we need an effective tunneling barrier height less than 120 meV. This can be achieved using a Type II heterostructure. Unfortunately, a small effective band gap requires a steep band edge density of states or else the current will pass through the band tail states and never see the barrier. As we will see in the next section, there are states below the band edge that prevent the tunneling junction from fully turning off. At that point the switching becomes controlled by the energy filtering mechanism. Consequently, modulating the thickness of the tunneling barrier will never give a steep subthreshold swing at high current densities!

4.3   Energy Filtering Switching Mechanism

It is possible use energy filtering as a switching mechanism. This is also called density of states switching. The energy filtering switch is illustrated in Fig. 4.1(c-d). If the conduction and valence band don't overlap, no current can flow. Once they do overlap, current can flow. An ideal density of states switch would be designed to switch abruptly from zero-conductance to the desired on conductance when the conduction and valance band overlap, thus displaying zero subthreshold swing voltage [13]. Unfortunately, the band edges are not perfectly sharp and so there is a finite density of states extending into the band gap.

In order to determine how steep the energy filter switching mechanism is, we need to determine how steep the electronic band edges are. While science has good knowledge on the magnitude of semiconductor bandgaps, there is not much information regarding the sharpness of the band edges. Although there are no good direct measurements of the band edge density of states, we can infer it from optical[14] and electronic measurements[15].

Typically the band edge density of states falls off exponentially below the band edge. We can parameterize this fall off with the term S_{DOS} which represents how many millivolts you need to go below the band edge to reduce the density of states by a decade. Below the bandgap, the optical absorption coefficient also falls of exponentially and is called the Urbach tail[14]. In intrinsic GaAs, the absorption falls off at 17meV/decade[16]. In intrinsic Si absorption falls off at 23mV/decade[17]. For electrons, we can hope to see a similar limit on the band edge steepness, S_{DOS}. This may seem promising, but such a steep result has not been vindicated by electrical transport measurements.

Electrically measured joint density of states have generally indicated a steepness >90mV/decade, unlike the intrinsic optical Urbach measurements which are <60mV/decade in good semiconductors. We attribute this broadening to the spatial inhomogeneity and on heavy doping that appears in real devices. Effectively, there are many distinct channel thresholds in a macroscopic device, leading to threshold broadening. Fortunately, this can be ameliorated. We can see this from the optical absorption in doped GaAs. When GaAs is doped with Si to 2×10^{18}/cm^3 the absorption falls off at a rate of 30meV/decade[16]. If the doping is further increased to 10^{20}/cm^3, the absorption falls of at a rate worse than 60meV/decade[18]. This means that if a tunnel switch is heavily doped, it will be unable to employ the density of states energy filtering mechanism to achieve a subthreshold swing voltage smaller than 60 mV/decade! Furthermore, in the doped optical absorption measurements, the band edge density of states is reduced by the free carriers screening potential fluctuations[19, 20]. Unfortunately, in the depletion region of a TFET, there are no free carriers to screen the potential variations and so that the band tails will be even worse in electronic devices.

4.3.1   Minimum Effective Band Gap

In addition to limiting the subthreshold swing voltage, the band edge density of states can limit the on/off ratio if the effective band gap (tunneling barrier height) is too small. If we want a particular on-off ratio, the barrier height in the off-state, $E_{g,eff}$, must be large enough to suppress the band edge density of states, S_{DOS}, by that on-off ratio. Consequently, we get the following limit:

$$E_{g,eff} \geq S_{DOS} \times \log(I_{on}/I_{off}) \tag{4.3.1}$$

For instance, if we want to use tunneling barrier width modulation we need a barrier height less than 120 mV. For six decades of on-off ratio, S_{DOS} must be steeper than 120/6 = 20 mV/decade which has not yet been achieved. Furthermore, the steepest turn-on will come from band edge rather than the tunneling barrier thickness modulation if we had S_{DOS} = 20 mV/decade. Consequently, modulating the thickness of the tunneling barrier will never give a steep subthreshold swing at high current densities.

4.4   Measuring the Electronic Transport Band Edge Steepness

To interpret electrical transport measurements, we need to look at the absolute conductance, I/V versus bias voltage V in a tunneling diode. The absolute conductance is proportional to the tunneling joint density of states. This is discussed in detail in [15]. Investigating the electronic steepness in a two terminal pn-junction measurement allows the band alignment to be controlled directly, without concern for gate efficiency. Both barrier thickness modulation and density of states switching changes the resistance of the tunneling junction. Consequently, we need to measure the change in resistance or conductance with bias rather than the change in current with bias. This can be seen from following model for the tunneling current:

$$I \propto \int (f_C - f_V) \times \top \times D_J(E) \times \partial E \tag{4.4.1}$$

(f_C-f_V) is the difference between the Fermi occupation probabilities on the p and n sides. $\top$ is the tunneling probability across the junction and $D_J(E)$ is the joint density of states between the valence band on the p-side and the conduction band on the n-side. We are interested in measuring the voltage dependence of the tunneling joint density of states, $\top \times D_J(E)$, in the integrand of Eq. (4.4.1). Since $\int (f_C - f_V) \times \partial E = qV$, dividing the current by the voltage approximately eliminates the effect of the Fermi levels[15]. In a three terminal transistor measurement the source drain bias would control the Fermi levels while the gate bias would control $\top \times D_J(E)$. This allows us to use a two terminal current voltage measurement to determine the joint density of states of a tunnel junction. Consequently, a two terminal source drain measurement in a TFET can be used to interpret the steepness of the tunnel joint density of states, without being limited by the gate efficiency. We will now interpret some specific cases from the experimental literature by plotting I/V versus V:

First we consider the 2-terminal current-voltage characteristics of an InAs/AlSb/Al$_{0.12}$Ga$_{0.88}$As heterojunction backward diode [21]. In this diode the tunneling barrier thickness is fixed by the AlSb thickness and so the tunneling is entirely due to the density of states overlapping. The I-V curves are in Fig. 4.2(a) and the absolute conductance is in Fig. 4.2(b). As seen in Fig. 4.2(a), the current diverges on a semilog plot at V=0, preventing direct interpretation. Likewise, the differential conductance diverges on a semilog plot at the Esaki peak. Thus a current or a differential conductance plot does not give us the information we want. By contrast, in Fig. 4.2(b), the absolute conductance, I/V, smoothly varies from reverse bias, through the origin, to forward bias.

The conductance is proportional to tunneling joint density of states, which can be parameterized by the inverse of the semilog slope of the conductance, called the semilog conductance swing voltage. This is equivalent to the steepness of the tunneling joint density of states in mV/decade shown by the inverse slope of the diagonal line in Fig. 4.2(b). In the figure, the semilog swing voltage of the absolute conductance is 98mV/decade, and it measures the tunneling joint density of states. It has one of the steepest experimentally measured tunneling joint density of states .This is likely due to the Type III band alignment permitting low doping levels ~1.4×10^{17}/cm^3 near the junction region.

In Fig. 4.2(c), we consider a germanium backward diode [22]. This has the steepest semilog conductance swing voltage of 92 mV/decade that we could find in the literature. Next, In Fig. 4.2(d), we show the current and conductance for an InAs homojunction diode at two different doping levels (N$_A$ = 1.8x10^{19}; N$_D$ = 3x10^{18} and 1x10^{19}) [23]. When the n-side doping is decreased from 1×10^{19}/cm^3 to 3×10^{18}/cm^3 the absolute conductance I/V swing improves from 570mV to only 180mV/decade falling far short of our goal. This clearly illustrates that while smearing the band-edge by doping is very bad, but that even the lower doped samples perform poorly.

We can apply the same 2-terminal analysis to 3-teminal TFETs. In a TFET we can either make a 2-terminal source-drain measurement, or a 3-terminal I$_D$-V$_G$ measurement. In a 3-terminal measurement, the subthreshold swing voltage will not give the tunneling joint density of states, since some voltage is lost across the gate oxide.

We avoid gate issues by doing a 2-terminal measurement. If the critical tunneling junction is at the source-channel junction, we need to fix the V_{Gate}-V_{Drain} voltage while measuring source/drain current versus source/drain voltage (I_D-V_S). Since we want to measure the band edge mechanism without being confounded by the gate modulation mechanism, we leave the V_{Gate}-V_{Drain} voltage fixed. As the gate potential will have a strong influence on the channel potential, the drain will not be able to effectively control the source-channel junction and so it is best to vary V_{GS}. If on the other hand the critical tunneling is at the channel/drain junction we need to fix the V_{Gate}-V_{Source} voltage while measuring source/drain current versus source/drain voltage (I_D-V_D).

In Fig. 4.2(e), we fix the V_{Gate}-V_{Source} voltage while measuring source/drain current versus source/drain voltage (I_D-V_D) of an $In_{0.53}Ga_{0.47}As$ TFET, that has a poor gate oxide[24]. Effectively, this is a 2-terminal measurement on a 3-terminal device. The semilog conductance swing voltage is 165mV/decade. The corresponding 3-terminal measurement shows a worse subthreshold swing voltage ~216 mV/decade. Since the two terminal measurement is not limited by the gate oxide, it reflects the junction's intrinsic tunneling properties. This shows the value of a proper 2-terminal measurement to analyze a TFETs potential performance when the gate oxide has poor quality.

In TFETs, subthreshold swing voltages less than 60 mV/decade have been measured, but only at extremely low current densities of ~1nA/µm. In Backward diodes and Esaki diodes the low current densities have been obscured by trap assisted tunneling and forward leakage current. Moreover, at current densities measured in tunneling diodes, tunneling barrier width modulation is weak, and I/V versus V reflects the band edge density of states.

Measuring the steepness of the conductance, I/V, in mV/decade of a tunneling diode, or of a TFET source-drain I-V, will give the tunneling joint density of states. This tells us the potential subthreshold swing voltage that we can expect from a TFET based on that tunneling junction at reasonable current densities. Looking at the best tunneling diodes to date, we find that they all have a semilog conductance swing voltage worse than 60 mV/decade. This is because they are macroscopic devices with considerable threshold inhomogeneity leading to multiple channels, each with a different threshold, smearing the subthreshold swing voltage. In the next section we suggest some remedies.
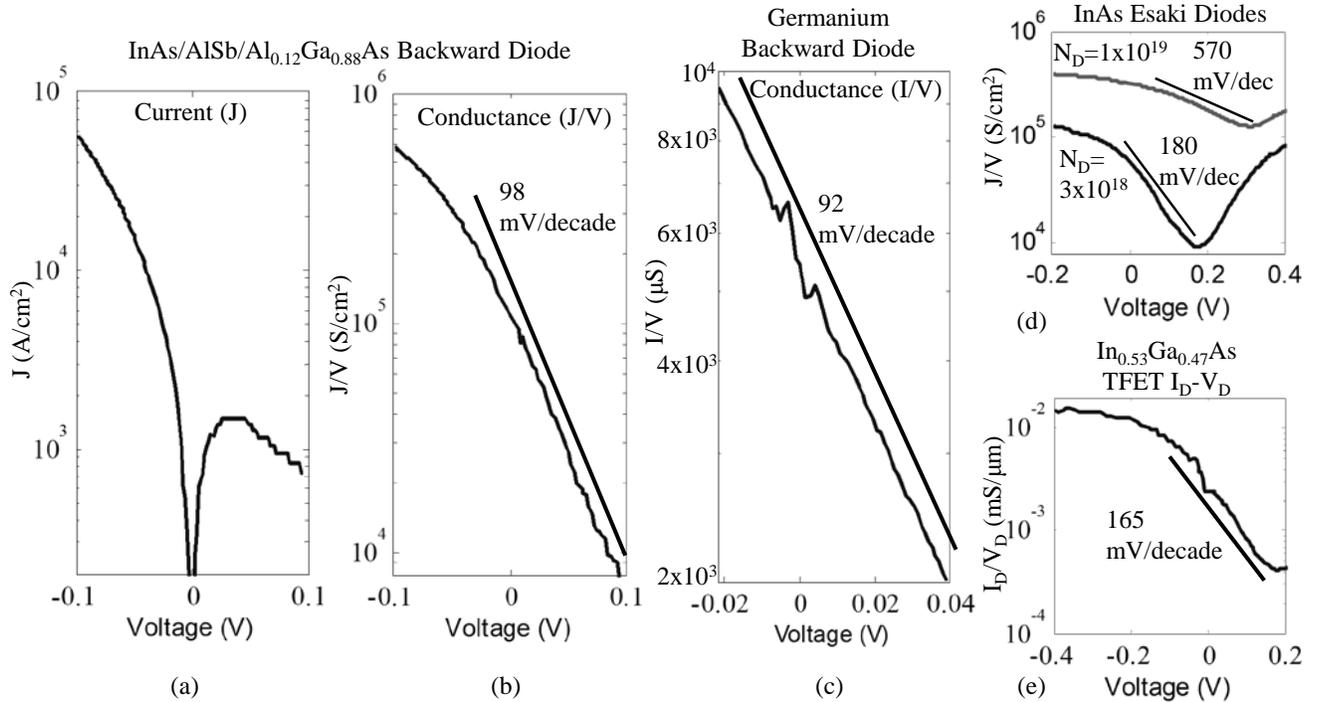
Fig. 4.2: (a-b) The current and conductance for an InAs/AlSb/Al$_{0.12}$Ga$_{0.88}$As heterojunction diode [21], (c) a Germanium diode [22], and (d) InAs diodes [23] are plotted. At V=0, the current diverges on a log plot and so the logarithmic slope is meaningless. Fortunately, the conductance is proportional to the tunneling density of states. (e) The G=I$_D$/V$_D$ versus V$_D$ for an In$_{53}$Ga$_{0.47}$As TFET is plotted[24]. The measured subthreshold swing voltage is 216 mV/decade while the semilog conductance swing voltage is 165 mV/decade. Since the I$_D$-V$_D$ characteristic is not limited by the gate oxide, it reflects the junction's steeper intrinsic tunneling properties.

4.5   Correcting Spatial Inhomogeneity

So far the prospects of designing a steep tunneling junction at high current densities seems quite bleak. Modulating the thickness of the tunneling barrier will not work and a band edge density of states steeper than 60 mV/decade has not been measured in electrical transport. In order to get better performance, novel geometries that provide spatial homogeneity,

eliminate doping, and promote atomic perfection are needed. Modulation doping that moves the dopants away from the tunneling junction may help, but the ideal geometry would use electrostatic doping through gates. This can be achieved through the bilayer[10, 11] structure in Fig. 4.3(a) or alternatively, one could try to make lateral double gate structure as shown in Fig. 4.3(b). Additional structures that remove doping from the tunneling junction and preserve the material quality are needed.

In addition to eliminating the doping, we need to eliminate any other sources of spatial inhomogeneity. This can come from rough heterojunctions, atomic thickness fluctuations, or any other non-ideality. Making small devices that encompass a single quantum wavefunction will still be inhomogeneous from device to device, but a single device will be more likely to show the intrinsic energy sharpness which has not yet been measured by electrical transport.

Alternately, the electrical transport measurements can be performed at low temperature, sharpening up the individual energies, and providing an opportunity to measure the discrete levels of the inhomogeneous distribution. Such a low temperature device would not be practical for use as a switch, but would provide scientific information about the inhomogeneities.

Part of the inhomogeneity arises from thickness fluctuations in conventional quantum wells. There have now emerged monolayer semiconductors such as MoS_2 that can precisely define the layer thickness, hopefully eliminating the problems of spatial homogeneity.
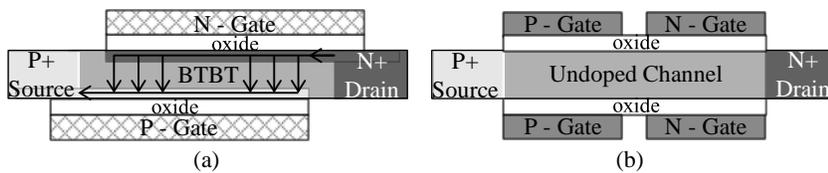
Fig. 4.3 (a) A bilayer TFET is shown. By applying opposite biases to both the N and P gates, both an electron and a hole channel form in the undoped channel, allowing band to band tunneling as shown. (b) A lateral TFET with both P and N gates is shown. The tunneling junction at the center of the channel is formed electrostatically without dopants. By eliminating doping in both stuctures, a steep density of states can be achieved.

4.6   PN-Junction Dimensionality

To further improve the performance of a tunneling junction, we need to maximize the on-state conductance and minimize the overdrive voltage. (The overdrive voltage is the extra voltage needed beyond the subthreshold regime to get the desired conductance.) This is strongly dependent on the actual geometry of the tunneling junction. Fortunately, confining the carriers in the tunneling direction provides four benefits that help achieve this [25].

1. The carrier velocity is increased and set by the confinement energy.
2. A higher electron energy can increase the tunneling probability.
3. Shrinking the region in which the electron is allowed will cause a greater percentage of the electron density to be in the barrier and thus the tunneling wave-function overlap increases.
4. Reducing the dimensionality results in a sharper density of states which reduces the overdrive voltage needed to get the full conductance.

Whenever specifying a pn junction it is also necessary to specify the dimensionalities of the respective p, and n regions. In Fig. 4.4 we show nine different possible pn junction dimensional combinations. In the following sections we analyze each of these devices and ask which are the most promising for adaptation into a TFET?
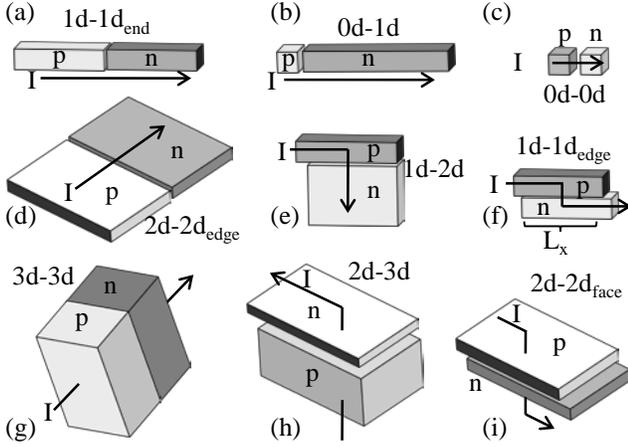
Fig. 4.4: We identify nine distinct dimensionality possibilities that can exist in pn junctions. Each of the different tunneling pn junction dimensionalities shown have different turn on characteristics.

4.6.1   1d-1d$_{end}$ Junction

A 1d-1d$_{end}$ pn junction describes tunneling within a nanowire [26] or carbon nanotube [27] junction as schematically represented in Fig. 4.4(a). Tunneling is occurring from the valence band on the p-side to the conduction band on the n-side. For a transistor, the gate is not shown as there are many possible gate geometries. The band diagram across this junction is given by Fig. 4.5(a).

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 voltage (k_bT/q or less) will result in a large change in the density of states but only a small change in the tunneling barrier thickness. Consequently, we assume that the tunneling probability is roughly a constant, $\top$, and will not change significantly for small changes in the control voltage. Initially, we also assume the tunneling probability is independent of energy and can be given by an energy averaged tunneling probability. We will discuss the energy dependence $\top(E)$ in the next section. The tunneling probability, $\top$, is the probability that an electron in a given mode tunnels through the barrier and end up on the other side. It is often given by a WKB approximation: $\top = \exp\left(\int kdx\right)$.

We also define $V_{OL}=qE_{OL}$ to be the overlap voltage between the conduction and valence bands as shown in Fig. 4.5(a). 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}. Making these approximations allows us to focus on the effects of changing the dimensionality and discover some new insights into tunneling in reduced dimensionality systems.

The 1d-1d$_{end}$ 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 Fig. 4.5(b). The current flow is controlled by the difference in the Fermi levels, which is V_{SD}, as shown. Current is given by charge × velocity × density of 1d states. Since the energy dependence of the velocity and 1d density of states exactly cancel, we get the quantum of conductance: $I=(2q^2/h)\times V_{SD}\times\top$.

Now to properly consider the transition from conduction band to valence band, we look at the band diagram given in Fig. 4.5(a). Initially, we consider the situation 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 Fig. 4.5(a), 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. Consequently, it is V_{OL} and not V_{SD} that controls the current:

$$I_{1d-1d} = \frac{2q^2}{h}\times V_{OL}\times\top \tag{4.6.1}$$
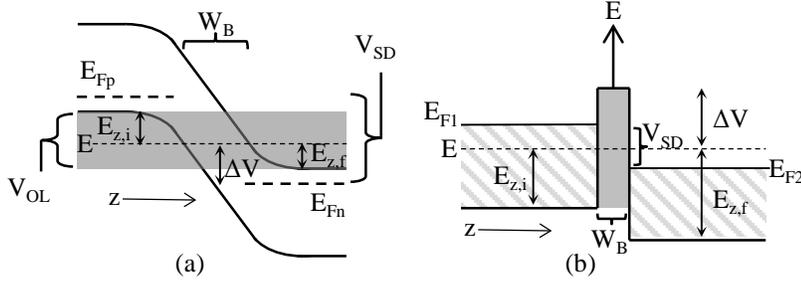
Fig. 4.5: (a) Energy band diagram for the tunnel pn junction showing that the relevant voltage is the overlap voltage and not the source drain voltage. (b) Energy band diagram for a typical 1d quantum of conductance showing that the relevant voltage is the source drain voltage.

4.6.1.1   Small Source Drain Bias Limit

Instead of assuming that there is a large bias across the tunneling junction, we can also consider the opposite limit where $V_{SD} < 4k_bT/q$. 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 optionally Taylor expand $f_c - f_v$:

$$f_{c,v} = \frac{1}{e^{(E-E_{Fc,v})/k_bT} + 1} \tag{4.6.2}$$

$$f_c - f_v \approx \frac{(E_{Fc} - E_{Fv})}{4k_bT} \approx \frac{qV_{SD}}{4k_bT} \tag{4.6.3}$$

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:

$$G_{1d-1d} = I_{1d-1d} \times \frac{q}{4k_bT} \tag{4.6.4a}$$

$$G_{1d-1d} = \frac{2q^2}{h} \times \top \times \frac{qV_{OL}}{4k_bT} \tag{4.6.4b}$$

This is true for all of the following devices to be considered in next sections 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. The exact integral over the Fermi functions is discussed in Chapter 6 for the 1d-1d$_{end}$ case.

Eq. (4.6.4b) illustrates a fundamental tradeoff between switching voltage and switch conductance. Even if the tunneling probability is 1, the conductance will be limited by $qV_{OL}/4k_bT$. For biases less than $4k_bT$, the conductance is reduced due to the thermal distribution of carriers. We can express this as a voltage-resistance product that will be limited to $>2hkT/q^3$. The voltage-resistance product says that low voltage switches inherently have high resistance, while high-conductivity switches will also require high voltage.

4.6.1.2   Fermi's Golden Rule Derivation

The current can be derived in a different manner using the transfer Hamiltonian method [25, 28-31]. 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[31]. It was then expanded by Bardeen[28] for tunneling in superconductors and then the case of independent electrons was considered by Harrison[30]. The transfer Hamiltonian method is just an application of Fermi's golden rule with a clever choice of states and perturbing Hamiltonian. The current density is given by Fermi's golden rule:

$$J = 2q \times \frac{2\pi}{\hbar} \sum_{k_i,k_f} |M_{fi}|^2 \delta(E_i - E_f)(f_C - f_V) \tag{4.6.5}$$

The calculation of the matrix element M_{fi} is in done in [25] and [30] and is given by:

$$M_{fi} = \frac{\hbar^2}{2m} \sqrt{\frac{k_{Z,f} k_{Z,i}}{L_{Z,f} L_{Z,i}}} \times \sqrt{\top} \times \delta_{k_{X,i}, k_{X,f}} \delta_{k_{Y,i}, k_{Y,f}} \tag{4.6.6}$$

In this equation, $k_{\alpha,i}$ and $k_{\alpha,f}$ are the α-component of the wave-vector in the initial and final states respectively. $L_{Z,i}$ and $L_{Z,f}$ are the lengths along the tunneling direction of the initial and final sides of the junction. Using this method allows us to extend the transfer Hamiltonian approach to the trickier reduced dimensionality cases by simply summing over fewer states. When quantum confinement is used in the tunneling direction, two effects will result in a large matrix element and thus a higher conductance. First, k_Z will be set to a large value corresponding to the increased velocity due to confinement. Second, L_Z will also be shorter. By shrinking the region in which the electron is allowed, a greater percentage of the electron density is in the barrier and thus the tunneling wave-function overlap increases.

4.6.2 Energy Dependent Tunneling Probability

A significant energy dependence arises at low energies where the WKB approximation breaks down. The tunneling probability approaches zero as the energy approaches zero. At small energies relative to the barrier height, the wave function begins to approach infinite barrier boundary conditions, where it is almost zero amplitude at the barrier. Therefore the tunneling probability has to approach zero at low energy.

The 1 band tunneling probability through a rectangular barrier, as shown in Fig. 4.5(b), can be found be matching boundary conditions using propagation matrices[32], is given by:

$$\top = \frac{1}{\dfrac{\left(\sqrt{E_{z,i}} + \sqrt{E_{z,f}}\right)^2}{4\sqrt{E_{z,i}E_{z,f}}} + \dfrac{\left(E_{z,i} + \Delta V\right)\left(E_{z,f} + \Delta V\right)}{4\Delta V \sqrt{E_{z,i}E_{z,f}}} \sinh^2(\kappa W_B)} \tag{4.6.7}$$

We considered the situation where the initial and final energy, $E_{z,i}$ and $E_{z,f}$ respectively, are different as shown in Fig. 4.5(b). The barrier height relative to the tunneling energy, E, is given by ΔV. The barrier width is W_B. The wavevector in the tunneling barrier is given by: $\kappa = \sqrt{2m\Delta V}/\hbar$. For a typical barrier the sinh term will be large and so we get:

$$\top \approx \frac{16\Delta V \sqrt{E_{z,i}E_{z,f}}}{\left(E_{z,i} + \Delta V\right)\left(E_{z,f} + \Delta V\right)} \exp\left(-2\kappa W_B\right) \tag{4.6.8}$$

At small energies, $E \ll \Delta V$, we get:

$$\top \approx \frac{16\sqrt{E_{z,i}E_{z,f}}}{\Delta V} \exp\left(-2\kappa W_B\right) \tag{4.6.9}$$

Thus we see that there is an energy dependent pre-factor to the WKB exponential. As the energy goes to zero, the tunneling probability goes to zero.

Since the exact form of the tunneling probability will be dependent on the barrier shape, we will continue to assume that an averaged tunneling probability, $\langle \top(E) \rangle$, can be used when calculating the current in the following sections. This still captures the key voltage dependence. The main result to remember is that at small overlap voltages the initial turn-on will be limited by the tunneling turn on.

In the band-to-band tunneling case, the probability will still be given by Eq. (4.6.8) if we assume a rectangular barrier. The only change is that $E_{z,i}$ is the hole energy and $E_{z,f}$ is the electron energy as shown in Fig. 4.5(a). This can be found by computing the tunneling matrix element used in Eq. (4.6.5). V_{OL} represents the available kinetic energy. When there is no confinement in the tunneling direction, such as the 1d-1d$_{end}$ case, the tunneling probability, Eq. (4.6.9), will be maximized when $E_{z,i} = E_{z,f} = qV_{OL}/2$:

$$\top_{max} \approx \frac{8V_{OL}}{\Delta V} \exp\left(-2\kappa W_B\right) \tag{4.6.10}$$

This means that the tunneling probability will be linearly proportional to V_{OL} at turn-on and a finite V_{OL} of $\Delta V/8$ is needed for the pre-factor to reach 1. Once the prefactor is 1, we assume the WKB approximation is valid.

There is also a secondary energy dependence that affects the tunneling probability. In the 3-dimensional WKB approximation, a large transverse energy will reduce the tunneling probability. Since the transverse energy is limited by the available overlap voltage, V_{OL}, at threshold the tunneling problem becomes more 1-dimensional, and the transverse energy can be neglected. The impact of accounting for the transverse energy on the 1d-1d$_{end}$ 2d-2d$_{edge}$ and 3d-3d cases is discussed in Chapter 6

4.6.3   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 is shown in Fig. 4.4(g).

The 3d 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:

$$\partial I = N_{\perp \, states} \times \frac{2q}{h} \times \langle \top \rangle \times \partial E \tag{4.6.11}$$

The number of transverse states is the number of k-states within the maximum transverse energy at a given energy and is given by the number of 2d states: $N_{\perp} = (AmE)/(2\pi\hbar^2)$. A is the area of the tunneling junction. The transverse energy is limited by the closest band edge and peaks in the middle of the overlap. Integrating Eq. (4.6.11) gives:

$$I_{3d-3d} = \frac{1}{2}\left(\frac{Am^*}{2\pi\hbar^2} \times \frac{qV_{OL}}{2} \right) \times \frac{2q^2}{h} V_{OL} \times \langle \top \rangle \tag{4.6.12}$$

$$=\text{No. of 2d Channels} \times \text{1d Conductance}$$

This is the same as taking the appropriate limits of Kane's tunneling theory [7] (except for a factor of $\pi^2/9$).

4.6.4   2d-2d$_{edge}$ Junction

A 2d-2d$_{edge}$ junction is shown in Fig. 4.4(d). 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:

$$I_{2d-2d,edge} = \frac{2}{3}\left(\frac{L_X \sqrt{m^*}}{\pi\hbar} \times \sqrt{qV_{OL}} \right) \times \left(\frac{2q^2}{h} \times V_{OL} \times \langle \top \rangle \right) \tag{4.6.13}$$

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

Where L_X is the length of the junction.

4.6.5   0d-1d Junction

A 0d to 1d junction represents tunneling from a quantum dot to a nanowire as shown in Fig. 4.4(b). 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. We analyze this junction as building block for the 2d-3d and 1d-2d junctions. To build a real 0d-1d device, we also need to electrically contact the quantum dot. Therefore we consider a more realistic situation that includes this. This becomes a single electron transistor (SET) as shown in Fig. 4.6.

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 [33], 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 $R=(1/\tau)\times\langle\top\rangle$. This can be converted to a current by multiplying by the electron charge, and a factor 2 for spin to give:

$$I = \frac{4q}{h} \times E_Z \times \langle \top \rangle \tag{4.6.14}$$

This is the same result that one obtains from the transfer Hamiltonian method.

To include coupling into the dot, we add a second nanowire to supply current, as shown in Fig. 4.6 and form a "single electron transistor" [34]. We assume that the second nanowire has the same tunneling probability/coupling strength to the quantum dot as the original one. Unlike a conventional SET, we want the current to be high enough and the dot be large enough that we do not see any coulomb blockade effects. The tunneling event out of the dot follows sequentially after tunneling in. Consequently, the current is cut in half:

$$I_{0d\text{-}1d} = \frac{2q}{h} \times E_Z \times \langle \top \rangle \qquad (4.6.15)$$

As seen in Fig. 4.6(e), the tunneling occurs at a single energy and will result in a sharp turn on once the bands overlap. This is one of the key benefits of quantum confinement. The current density is concentrated in a narrow energy range which allows for a smaller V_{OL}. The can be contrasted with the 1d-1d$_{end}$ case, Eq. (4.6.1), where the current flows over the entire energy range corresponding qV_{OL}. The width of the 0d-1d energy range will be given by the broadening of the energy level in the quantum dot. 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 causes a significant amount of broadening. Each contact will broaden the level by γ_0 for a total broadening of $2\gamma_0$[35]:

$$\gamma_0 = \frac{\hbar}{\tau} = \frac{1}{\pi} \times E_Z \times \langle \top \rangle \qquad (4.6.16)$$

In the limit that $\top \to 1$, the 0d-1d case will become the 1d-1d$_{end}$ case with a perfect quantum of conductance: $I = 2q^2/h \times V_{OL}$. However, in a realistic situation $\top << 1$, and so we can use quantum confinement to concentrate the current at a single energy and significantly reduce V_{OL}.

The quantum confinement also has an added benefit of increasing the tunneling probability itself. This can be seen from Eq. (4.6.9) In the 1d-1d$_{end}$ case, $E_{z,i}$ and $E_{z,f}$ are both limited by V_{OL}. In the 0d-1d case, only $E_{z,f}$ will be limited by V_{OL}. $E_{z,i}$ can be set to a large value by the quantum confinement.

(a) (b) (c) (d) (e)
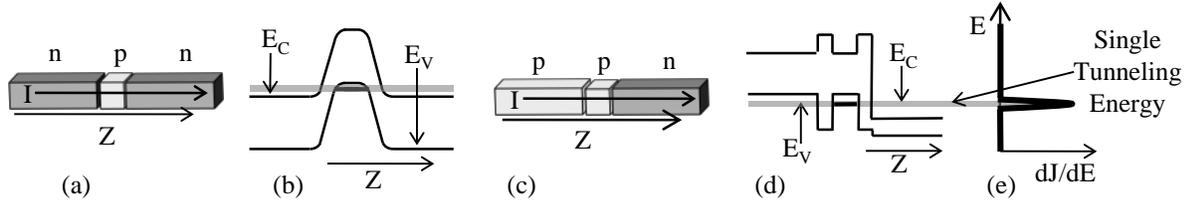
Fig. 4.6: (a) A 0d-1d junction converted into a more realistic 1d single electron transistor (SET) structure. (b) Band diagram corresponding to the SET. (c) An alternate SET structure with a p-type contact (d) Band diagram corresponding to the alternate SET (e) All the current is concentrated around a single energy, which allows for a small overlap voltage V_{OL}, and thus a small overdrive voltage V_{OV}.

4.6.6   2d-3d Junction

A 2d-3d tunneling junction is typical in vertical tunneling junctions where the tunneling occurs from the bulk to a thin confined layer [3]. The thin layer can either be a thin inversion layer or a physically separate material [3, 36, 37]. A generalized schematic of this tunnel junction is shown in Fig. 4.4(h).

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. The 1d problem is better described by tunneling from a quantum dot to a nanowire as shown in Fig. 4.4(b).

To find the 2d-3d current, we simply multiply the 0d-1d result, Eq. (4.6.14), by the number of 2d channels to get a current of:

$$I_{2d-3d} = \text{No. of 2d channels} \times \text{1d field ionization}$$

$$I_{2d-3d} = \left(\frac{Am}{2\pi\hbar^2} \times \frac{qV_{OL}}{2} \right) \times \left(\frac{4q}{h} \times E_Z \times \langle \top \rangle \right) \qquad (4.6.17)$$

Here, E_Z is the confinement energy of the 2d layer. We only included transverse states for a transverse energy up to $qV_{OL}/2$. For transverse energies larger than that, no state exists on both sides of the junction with the same total energy. This is the exact same result that comes from the transfer Hamiltonian method. Compared to the bulk 3d-3d case, confining one side of the junction resulted in the replacement of qV_{OL} with $4E_Z$. The quantum confinement can also increase the tunneling probability by fixing the electron energy in the quantum well and thus $E_{z,i}$ in Eq. (4.6.9).

Current can flow in along the transverse direction as shown in Fig. 4.4(h). Other methods such as tunneling into the quantum well can also be considered for making electrical contact.

4.6.7   1d-2d Junction

A 1d-2d junction describes tunneling between the edge of a nanowire and a 2d sheet as shown in Fig. 4.4(e). The derivation for this case is almost identical 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:

$$I_{1d\text{-}2d} = \text{no. of 1d channels} \times \text{1d tunnel ionization}$$

$$I_{1d-2d} = \left(\frac{L_X}{\pi \hbar} \times \sqrt{qm^*V_{OL}} \right) \times \left(\frac{4q}{h} \times E_Z \times \langle \top \rangle \right) \tag{4.6.18}$$

Comparing to the 2d-2d edge overlap formula, confining one side of the junction resulted in the replacement of qV_{OL} with $3E_Z$ and an increased $\langle \top \rangle$ by fixing $E_{z,i}$ in Eq. (4.6.9).

4.6.8   0d-0d Junction

This case represents tunneling from a filled valence band quantum dot to an empty conduction band quantum dot. It is schematically represented in Fig. 4.4(c). 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. Consequently, we consider the structure in Fig. 4.7

Current will only flow when the confined energy levels in each dot are aligned. This can be seen from Fig. 4.7(d). The two dots will only couple if the density of states in each dot overlaps. This results in an I-V curve that resembles a delta-function as shown in Fig. 4.7(e). We can estimate the peak current by considering the coupling strength between each dot and its contact as well as the coupling between dots. For simplicity, we will assume that the dots and contacts are symmetric. The coupling strength or broadening due to each contact is given by Eq. (5.5.17):

$$\gamma_0 = \frac{\hbar}{\tau} = \frac{1}{\pi} \times E_Z \times \langle \top_{contact} \rangle \tag{4.6.19}$$

$T_{contact}$ represents the tunneling probability between the contact and a dot.

The coupling strength between each dot is the matrix element between the dots and is given by Eq. (4.6.6). Since we have a single level in each dot, we can simplify the matrix element by using $k_Z = \pi/L_Z$ and $E_Z = \hbar^2 k_Z^2 / 2m^*$:

$$\left| M_{f i,0d-0d} \right| = \frac{1}{\pi} \sqrt{E_{Z,i} \times E_{Z,f} \times \langle \top \rangle} \tag{4.6.20}$$

$\top$ is the single barrier tunneling probability between the two dots. In order to maximize the current we want all the coupling strengths to be equal: $|M_{fi}| = \gamma_0$. Since $|M_{fi}| \propto \sqrt{\langle \top \rangle}$ it is possible to design the central barrier to have $|M_{fi}| \vartriangleright \gamma_0$. Unfortunately, doing this will cause the dots to strongly couple and will result in a level splitting that reduces the current. Consequently, we want to design γ_0 to be large and then design $|M_{fi}| = \gamma_0$. This means that the tunneling rate through each barrier is the same and given by $\gamma_0 / \hbar$. Since we have a three step tunneling process the peak current will be given by:

$$I_{peak} \leq \frac{2q}{3\tau} = \frac{2\gamma_0}{3\hbar} = \frac{2}{3} \times \frac{2q}{h} E_Z \langle \top_{contact} \rangle \tag{4.6.21}$$

The width of the tunneling peak is given by the broadening of the confined level, $2\gamma_0$. Additional broadening mechanisms such as electron-phonon interactions can further broaden the turn on and reduce the peak current by smearing out the levels and reducing the coupling strength between the dots. As with the 0d-1d case, in the limit that $\top \rightarrow 1$, the 0d-0d case will become the 1d-1d$_{end}$ case with a perfect quantum of conductance: $I = 2q^2/h \times V_{OL}$. However, in a realistic situation $\top << 1$, and so we can use quantum confinement to concentrate the current at a single energy and significantly reduce V_{OL}.

The 0d-0d peak current is almost identical to the 0d-1d case from Eq. (4.6.15) with the exception of a factor of 2/3 as this is now a 3 step tunneling process instead of a 2 step process. The key difference arises when evaluating $\top_{contact}$. If we design the 0d-0d system as shown in Fig. 4.7(b) both the initial and final tunneling energy is non-zero and set by the quantum confinement. This means that the prefactor in the tunneling probability in Eq. (4.6.9) can be close to 1 and so an additional overlap voltage is not needed to increase the tunneling probability. Nevertheless, this also means that the subthreshold swing voltage will be determined only by the sharpness of the confined energy levels and not the band edge. Alternatively, we can design the 0d-0d system so that the nanowire band edges line up with the confined energy levels as shown in Fig. 4.7(c). In this case, the subthreshold swing voltage will be determined by the sharper of the band edge or the confined level. However, we will lose the increased tunneling probability as the energy in the nanowire will be low. Overall we see that by using the 0d-0d structure in Fig. 4.7(b) we may be able to have an advantage over 0d-1d through the increased tunneling probability. On the other hand, the delta-function like shape of the I-V curve shown in Fig. 4.7(e) may make it difficult to design a

conventional logic circuit.
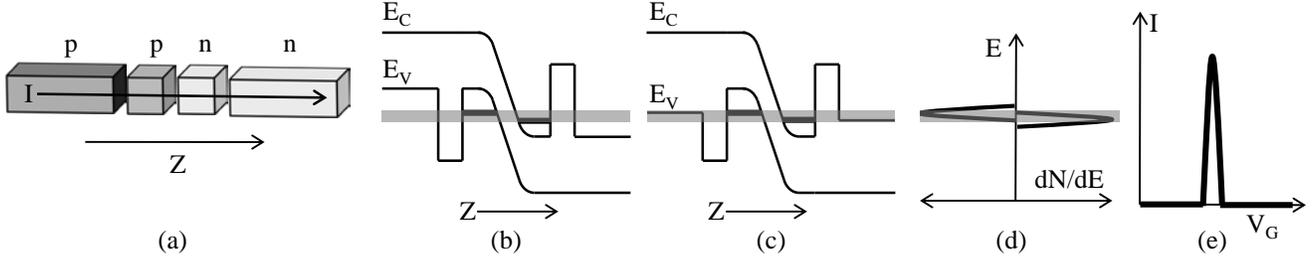
Fig. 4.7: The properties of a 0d-0d junction that is coupled to nanowire contacts are shown. (a) Schematic representation of the junction (b) Band diagram of the junction. Tunneling only occurs at a fixed energy when the two confined levels overlap. The fixed tunneling energy results in a higher tunneling probability. The subthreshold swing voltage will be determined by the sharpness of the levels. (c) Alternate band diagram that shows the nanowire band edges aligned with the confined levels. In this case, the subthreshold swing voltage will be determined by the sharper of the level or the nanowire band edge. Conversely, the tunneling probability will be lower as the energy in the nanowires is low. (d) Tunneling only occurs when the density of states in each dot is aligned. (e) The I-V curve resembles a delta function when the levels align.

4.6.9   2d-2d$_{face}$ Junction

A 2d-2d$_{face}$ junction describes tunneling from one quantum well to another through the face of the quantum well. This can be seen in resonant interband tunnel diodes [38-40]. The junction is schematically represented in Fig. 4.4(i). This is one of the most interesting cases as it is the closest to a step function turn-on.

The step function turn on can be seen by considering the conservation of transverse momentum and total energy. This depicted in Fig. 4.8(a). The lower paraboloid represents all of the available states in k-space on the p side of the junction and the upper paraboloid represents the available k-space states on the n 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 Fig. 4.8(a), 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 Fig. 4.8(b).

The current can be computed by using Fermi's golden rule. Due to the conservation of transverse momentum, every initial state is coupled to only one final state. Current can flow into each quantum state, along the quantum well, or through the face of the quantum well. We simply need to sum Eq. (4.6.5) over all initial states or final states:

$$I = 2q \times \frac{2\pi}{\hbar} \sum_{k_t} |M_{fi}|^2 \delta(E_C - E_V)(f_C - f_V) \tag{4.6.22}$$

Plugging in the 0d-0d matrix element, Eq. (4.6.20), converting the sum to an integral, and assuming a full valence band and empty conduction band gives:

$$I = 2q \times \frac{Am}{\pi^2 \hbar^3} \int E_{Z,i} \times E_{Z,f} \times \top \times \delta(E_i - E_f) dE_t \tag{4.6.23}$$

Finally evaluating the integral over the delta function gives an additional factor of ½ as $E_i - E_f = 2E_t - qV_{OL}$:

$$I_{2d-2d,face} = \frac{qmA}{\pi^2 \hbar^3} \times E_{Z,i} \times E_{Z,f} \times \langle \top \rangle \tag{4.6.24}$$

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$_{face}$ 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-2d$_{face}$ case has the same current as a 3d-3d case if $qV_{OL} = 2\sqrt{2}E_Z$. In practice E_Z can be much larger than qV_{OL}, providing the 2d-2d$_{face}$ case with a significant current boost. The quantum confinement also increases the tunneling probability itself as seen from Eq. (4.6.9). The pre-factor in the tunneling probability is no longer dependent on V_{OL} and is instead set to a large value by the quantum confinement.

Following the joint density of states, the current takes the form of a step function with respect to the gate voltage. This is because all of the tunneling current is concentrated near a single energy. 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, various broadening mechanisms will smear out the step-like turn-on function and this will be discussed later.
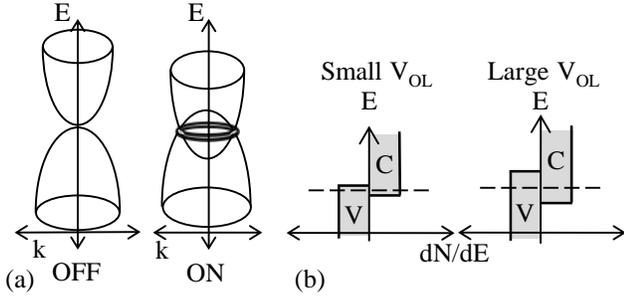
Fig. 4.8: Various characteristics of a 2d-2d$_{face}$ junction. (a) There is only a single tunneling energy because of the simultaneous conservation of energy and momentum. The energy versus wave vector paraboloids on each side of the junction only intersect at a single energy. (b) Even though the overlap 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.

4.6.10 1d-1d$_{edge}$ Junction

A 1d-1d$_{edge}$ junction represents two nanowires overlapping each other along the edge as shown in Fig. 4.4(f). This junction is similar to the 2d-2d$_{face}$ junction. The current can be found by summing over 1d set of transverse states in Eq. (4.6.22). The resulting current is:

$$I_{1d-1d,\,edge} = 2\frac{qL_X}{\pi^2\hbar^2}E_{Z,i}\times E_{Z,f}\times\sqrt{\frac{m}{qV_{OL}}}\times\langle\top\rangle \tag{4.6.25}$$

As in the 2d-2d$_{face}$ case the tunneling only occurs at a single energy due to the conservation of momentum and energy. 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 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 various broadening mechanisms will limit the peak conductance.

4.6.11 Tradeoff between Current, Device Size, and Level Broadening

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 when 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$_{edge}$ and 2d-2d$_{face}$ cases have a finite extent along the tunneling direction, restricting the normalization volume. This means that the tunnel interaction matrix element, $|M_{f\,i}|$, 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. The strong coupling will cause the interacting levels to repel each other and consequently limit the current. To prevent this and wash out the level repulsion, the level broadening, γ, needs to be greater than the level repulsion matrix element:

$$\gamma > \left|M_{fi}\right| = \frac{1}{\pi}\sqrt{E_{Z,i}\times E_{Z,f}\times\langle\top\rangle} \tag{4.6.26}$$

The broadening γ is typically caused by coupling to the contacts or by various scattering mechanisms. Unfortunately, this level broadening also smears out the sharp turn on of the 1d-1d$_{edge}$ and 2d-2d$_{face}$ junctions. Since the tunneling current is proportional to the matrix element, $|M_{f\,i}|$, there is a fundamental tradeoff between the level broadening and the tunneling current. The greater the on-state current, the more the level needs to be broadened to allow the electrons to escape into the contact. The 1d-1d$_{end}$, 2d-2d$_{edge}$ and 3d-3d junctions can be thought of as the limit where the levels are completely broadened into a continuous band. In the limit that $\top\rightarrow 1$, we cannot do better than a perfect quantum of conductance from the 1d-1d$_{end}$ 2d-2d$_{edge}$ and 3d-3d cases. However, any realistic device will have $\top << 1$, and so we can use the quantum confinement to increase the matrix element and to engineer the tradeoff between the broadening and the on-current to get a sharper turn on.

Another major broadening limit occurs for the 1d-2d, 2d-3d, 1d-1d$_{edge}$ and 2d-2d$_{face}$ cases when the transverse dimensions are reduced. Consider the 1d-1d$_{edge}$ case shown in Fig. 4.4(f). The current is flowing in along extended nanowires in the x-direction. At low energies, the wavelength along the x-direction will be very long and so only the tail of

the wavefunction will fit overlap region between the two quantum wires. In order to get a good transverse momentum matching at least half a wavelength needs to fit in the overlap region. Consequently, the turn on will be broadened by the energy corresponding to $\lambda = 2L_x$:

$$\gamma \approx \frac{\hbar^2 \pi^2}{2mL_x^2} = E_x \qquad\qquad (4.6.27)$$

The same limit applies for the 1d-2d, 2d-3d and 2d-2d$_{face}$ case. This means that these cases will lose their abrupt turn on if the transverse dimensions are too small. For extremely small dimensions 0d-1d or 0d-0d may be more favorable. Alternatively, tunneling contacts could be used, but a different broadening limit given by Eq. (4.6.19) will apply.

4.6.12 Comparing the Different Dimensionalities

Now that we have considered many different tunneling junction geometries, we've plotted a comparison of the different cases in Fig. 4.9. To plot the figures we used a reasonable tunneling probability of 1%. We assumed confinement energies of 130 meV, an effective mass of 0.1 and overlap lengths of 20 nm. There are four different broadening mechanisms that will limit the initial turn on, as indicated by the dotted lines. Using the above constants, the broadening mechanisms and the affected dimensionalities are summarized below:

- Transverse momentum matching: Eq. (4.6.27).
  - 1d-2d, 2d-3d, 1d-1d$_{edge}$, 2d-2d$_{face}$, ------------ γ=9.4 meV
- Matrix Element Broadening: Eq. (4.6.26).
  - 0d-0d, 1d-1d$_{edge}$ and 2d-2d$_{face}$ ----------------- γ=4.1 meV
- Contact Broadening: Eq. (4.6.19)
  - 0d-0d, and 0d-1d ------------------------------- γ=0.8 meV
- Tunneling Probability Turn-On: Eq. (4.6.9)
  - 1d-1d$_{end}$, 2d-2d$_{edge}$ and 3d-3d ------------------ γ=12.5 meV
  - 0d-1d, 1d-2d and 2d-3d ------------------------ γ=0.6 meV

To estimate the tunneling probability turn-on, we assume a barrier height of ΔV=100 meV, which assures a good on/off ratio as discussed in Section 4.3.1. The 1d-1d$_{end}$, 2d-2d$_{edge}$ and 3d-3d are unconfined in the tunneling direction and so we can set the prefactor in Eq. (4.6.10) equal to 1: *8V$_{OL}$/ΔV=1*. This gives V$_{OL}$=γ=12.5 meV. For the 0d-1d, 1d-2d and 2d-3d cases, one side of the junction is confined and so we need to use Eq. (4.6.9) with $E_{z,f} = qV_{OL}/2$ and set the prefactor in Eq. (4.6.9) equal to 1: $16\sqrt{E_{z,i}qV_{OL}/2}\big/\Delta V = 1$. This gives V$_{OL}$=$\gamma$=0.6 meV. The broadening due to contacts is twice the broadening from a single contact given by Eq. (4.6.19).

For each dimensionality, the largest form of broadening will dominate. For the 1d-2d, 2d-3d, 1d-1d$_{edge}$ and 2d-2d$_{face}$ junctions, the turn-on will be limited by the transverse momentum matching. The 1d-1d$_{end}$, 2d-2d$_{edge}$ and 3d-3d junctions are limited by the turn on of the tunneling probability. The 0d-1d and 0d-0d are limited by the contact broadening.

The turn-on conductance versus overlap control voltage V$_{OL}$ can be seen in Fig. 4.9 for all of the cases. The initial broadened turn on is represented by the dotted lines. For the 0d-0d case the entire line-shape is due to the broadening and is thus unknown, but the calculated width and height are still represented in the figure.

The nanowire based devices shown in Fig. 4.9(a) have the lowest conductance as they only tunnel at a single point. However, we see that introducing quantum confinement can still help increase the conductance when the tunneling probability is low. We also see that for the parameters chosen, the 0d-1d case captures all the benefits of the quantum confinement, while the 0d-0d case is a narrow pulse with a slightly lower peak conductance. In some situations, when the tunneling probability requires a larger voltage to turn on, the 0d-0d case can have a higher initial peak.

The edge tunneling devices shown in Fig. 4.9(b) have a higher conductance as the have a larger tunneling length. Consequently, we also normalize the current to the tunneling length. In these cases, maximizing the quantum confinement on both sides of the junction results in the highest conductance. The same applies for the area tunneling devices shown in Fig. 4.9(c).

Overall, we see that using quantum confinement in the tunneling direction can significantly increase the conductance and thus reduce the overdrive voltage when the tunneling probability is low.
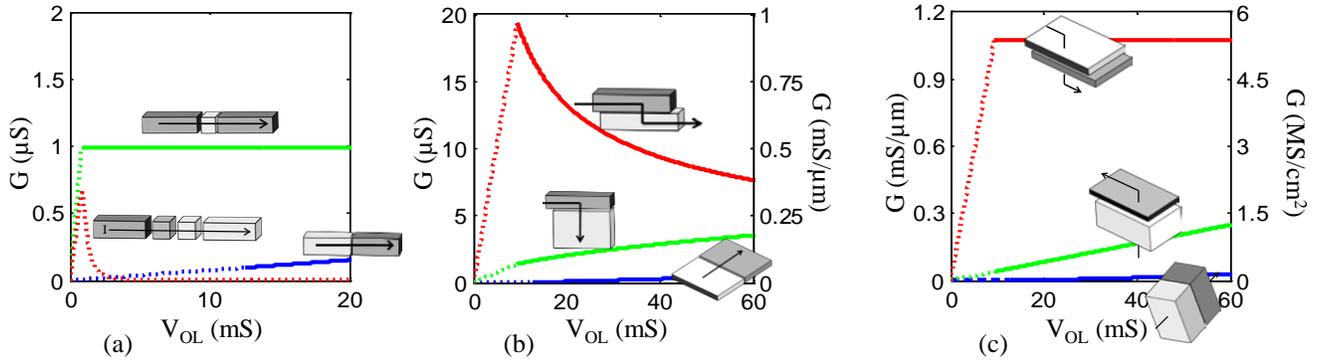
Fig. 4.9: The conductance curves for the different dimensionalities are plotted using the following parameters: $\langle \top \rangle$=1%, E_Z=130meV, L_X=20nm, m*=0.1m_e and ΔV=100meV. The dotted lines represent the initial broadened turn on where the lineshape is uncertain. (a) The 1d-1d$_{end}$, 0d-1d and 0d-0d cases are plotted. (b) The 2d-2d$_{edge}$, 1d-2d and 1d-1d$_{edge}$ cases are plotted. (c) The 3d-3d, 2d-3d and 2d-2d$_{face}$ cases are plotted.

4.7   Building a Full Tunneling Field Effect Transistor

Now that we have analyzed how to make a good tunneling junction we can consider what happens when we try to build a TFET. In considering the performance of a full TFET, one has to consider the subthreshold regime, gate efficiency, and any overdrive voltage needed to achieve the full on-state. While the voltage response is exponential over most of the drive range, it invariably saturates as the switch approaches the on-state. The overdrive voltage represents the extra voltage needed to achieve the full on-state after saturation has set in.

We can understand many of the design issues by considering three simple TFET structures that capture the essential physics behind most TFETs to date. We only analyze n-channel TFETs, where the gate modulates the n-type side of the pn-junction, since the analysis is almost identical for p-channel TFETs. Fig. 4.10(a) shows a double gate TFET with identical gates such that the tunneling occurs laterally at a single point at the source channel junction. The key design issues will be similar for other point tunneling devices such as nanowires or even single gate TFETs[12, 27]. Fig. 4.10(b) shows a vertical tunneling TFET where the gate overlaps the source. The gate inverts the source and a tunneling junction is formed within the source. A variety of schemes to optimize the vertical tunneling such as creating a doped pocket or using a vertical heterojunction have been tested, but they all operate on similar principles [3, 36, 37] and will face similar issues. Fig. 4.10(c) shows a newer device concept called the electron hole bilayer TFET [10, 11]. By applying opposite biases to both the N and P gates, both an electron and a hole channel form, allowing band to band tunneling as shown in Fig. 4.10(f). In all of the designs, the gate workfunction can be engineered to correctly set the threshold voltage.
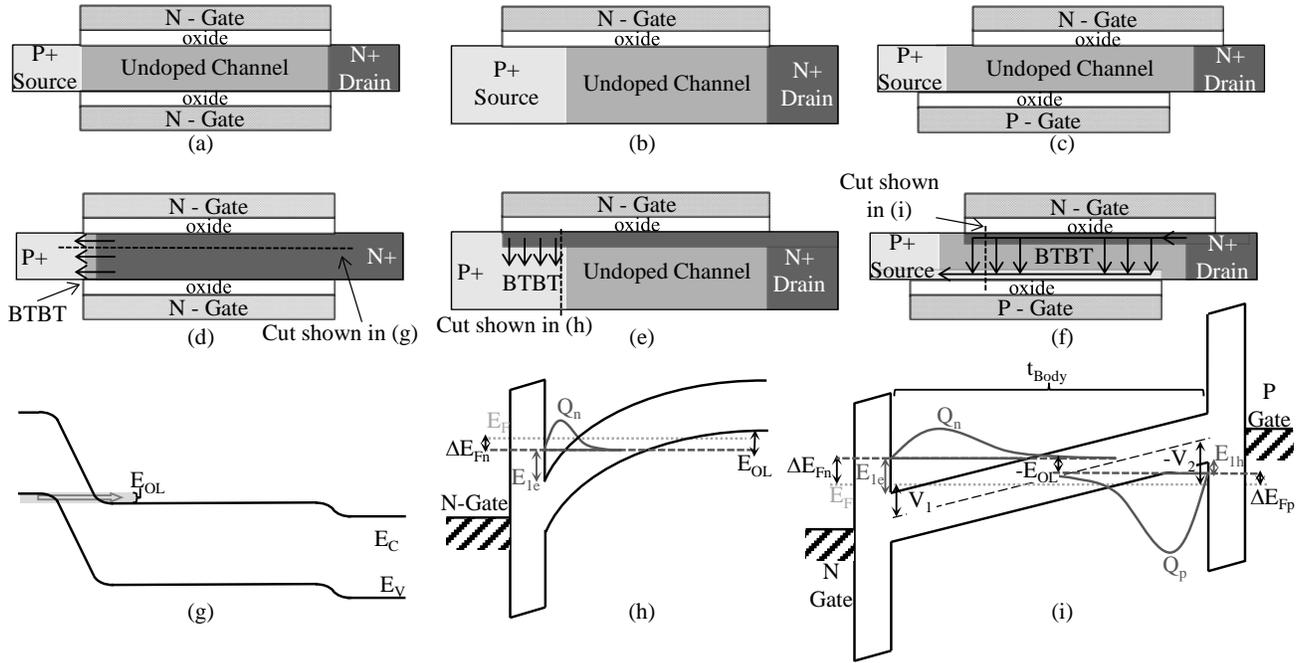
Fig. 4.10: Three representative tunneling FET designs are shown. (a) Lateral, (b) vertical and (c) bilayer TFETs are illustrated. The direction of tunneling for the lateral, vertical and bilayer is shown in (d), (e) and (f) respectively. The band diagrams parallel to the direction of tunneling is given in (g), (h) and (i) respectively.

4.7.1 Minimum Voltage Required

The operating voltage is given by:

$$V_{DD} = V_{OV} + SS \times Log\left(I_{on}/I_{off}\right) \tag{4.7.1}$$

Here we explicitly include the overdrive voltage, V_{OV}, and the subthreshold swing voltage, SS, as V_{OV} can be a significant part of the total voltage in a steep slope device. I_{on}/I_{off} is the ratio of the on and off state currents. V_{OV} is the extra voltage required after the conduction and valance band are aligned to get the desired on-state conductivity and is given by:

$$V_{OV} = \frac{dV_G}{dE_{OL}} \times \frac{1}{q} E_{OV} = \frac{1}{\eta_{gate}} \times \frac{1}{q} E_{OV} \tag{4.7.2}$$

E_{OL} is the overlap between the conduction and valence band edge as shown in Fig. 4.1(c) and 4.1(d). E_{OV} is the minimum energy overlap, E_{OL}, required to get the desired on-state conductivity. Typically, the conductance increases as E_{OL} increases as there are more states that contribute to the tunneling current and so a minimum energy overlap E_{OV} is needed. Fortunately, E_{OV} can be significantly reduced by confining the carriers in the tunneling direction as discussed in Section 4.6.

The next factor in computing V_{OV} is the gate efficiency, η_{gate}. Only a fraction of the voltage applied on the gate, V_G, contributes to changing E_{OL}. The gate efficiency can divided into two terms:

$$\eta_{gate} = \frac{1}{q}\frac{dE_{OL}}{dV_G} = \frac{1}{q}\frac{d\phi_s}{dV_G}\frac{dE_{OL}}{d\phi_s} = \eta_{el} \times \eta_{quant} \tag{4.7.3}$$

φ_S is the surface potential under the gate. The standard electrostatic gate efficiency (η_{el}) is just the change in the surface potential (φ_S) with respect to the gate bias (V_G):

$$\eta_{el} = d\phi_s/dV_G \tag{4.7.4}$$

There is also a quantum confinement efficiency (η_{quant}) for vertical and bilayer TFETs. In these structures a triangular quantum well is formed under the gate and the effective band edge is raised to the quantum level E_{1e} shown in Fig. 4.10(h) and Fig. 4.10(i). In the bilayer, both sides of the tunneling junction are confined as shown in Fig. 4.10(i). If the bias on the gate is increased, the triangular quantum well gets narrower and the energy level E_{1e} increases. This works against the gate bias reduces the change in the confined eigenstate energy overlap (E_{OL}). Consequently, we need to multiply by an additional quantum confinement efficiency:

$$\eta_{quant} = \left| \frac{1}{q} \frac{dE_{OL}}{d\phi_s} \right| \qquad (4.7.5)$$

So far we have assumed the tunneling process limits the current, but there is an additional requirement that channel have enough charge to conduct the current (i.e. the MOSFET in a TFET needs to be on as well) which is typically true in short channel TFETs. In a long channel device where a significant amount of charge is needed, the standard MOSFET electrostatics could dominate the overdrive voltage and needs to be accounted for. This would result in a very low η_{el} and increased E_{OV}.

4.7.2 Subthreshold Swing Voltage

The next step is to find the subthreshold swing voltage. The ideal TFET would rely upon a sharp band edge and would switch abruptly from zero-conductance to the desired on-conductance when the electron and hole eigenstate energies overlap. Unfortunately the band-edges are not perfectly sharp and thus there is a finite density of states extending into the band gap, smearing out the desired abrupt response. Conventional TFET modeling does not account for the smeared band edge density of states.

We consider the following simple model for the tunneling current in order to understand the various contributions to the Subthreshold Swing Voltage (SS):

$$I \propto \int (f_C - f_V) \times \top \times D_J(E) \times \partial E \qquad (4.7.6)$$

$(f_C\text{-}f_V)$ is the difference in the Fermi occupation probabilities of the conduction and valence bands, $\top$ is the transmission probability of a tunneling electron and $D_J(E)$ is the joint density of states. Below the band edge, E'_C, the density of states in the conduction band, $D_C(E)$, is given by:

$$D_C(E) = D_{C0} \times e^{-(E'_C - E)/qV_0}, E < E'_C \qquad (4.7.7)$$

Here D_{C0} is a constant prefactor for the electron density of states and E'_C is the electron eigenstate energy. We assume that the density of states falls off exponentially below the band edge with a semilog slope of V_0 and has a prefactor of D_{C0}. An exponential falloff is typical of band edges as seen in the optical absorption edge [18]. Above the band edge the density of states will simply be given by the 1d, 2d or 3d density of states depending on the device geometry. Similarly, the valence band edge density of states above the valence band edge, E_V', will be given by:

$$D_V(E) = D_{V0} \times e^{-(E-E'_V)/qV_0}, E > E'_V \qquad (4.7.8)$$

Here D_{V0}, is a constant prefactor for the hole density of states and E'_V is the hole eigenstate energy. For simplicity, we take the exponential slope, V_0, to be the same for conduction and valence band edges.

Now we consider the situation where the electron and hole eigenstates are not aligned as shown in Fig. 4.1(c). Ideally, no current would flow, but due to the band tails, an overlapping density of states exists as shown in Fig. 4.11. Combining the conduction and valence band density of states to get a joint density of states gives:

$$D_J(E) \propto e^{-|E_{OL}|/qV_0} \times \begin{cases} e^{-(E-E'_C)/qV_0}, & E \geq E'_C \\ 1, & E'_C > E > E'_V \\ e^{-(E'_V - E)/qV_0}, & E \leq E'_V \end{cases} \qquad (4.7.9)$$

E_{OL} is the overlap energy between the electron and hole eigenstates and is given by $E'_V\text{-}E'_C$ as shown in Fig. 4.1(c). Since the joint density of states has a maximum plateau in the bandgap region between E'_C and E'_V, we can approximate the current integral as:

$$I \propto \left(\int_{E'_V}^{E'_C} (f_C - f_V) \times \top \times \partial E \right) \times e^{-|E_{OL}|/qV_0} \qquad (4.7.10a)$$

$$I \propto I_0 \times e^{-|E_{OL}|/qV_0} \qquad (4.7.10b)$$

where the tunneling pre-factor is:

$$I_0 \equiv \int_{E'_V}^{E'_C} (f_C - f_V) \times \top \times \partial E \qquad (4.7.11)$$

Thus we have arrived at a simplified model for the tunneling current when band tails are present. Now we can compute the

subthreshold swing voltage (SS) using the definition:

$$SS \equiv dV_G / d\log(I) \tag{4.7.12}$$

Plugging Eq. (4.7.10) into Eq. (4.7.12) gives:

$$SS = \left(\frac{d\phi_S}{dV_G} \times \frac{d\log(I_0)}{d\phi_S} \quad + \quad \frac{d\phi_S}{dV_G} \frac{dE_{OL}}{d\phi_S} \times \frac{d\log\left(e^{-|E_{OL}|/qV_0}\right)}{dE_{OL}} \right)^{-1} \tag{4.7.13}$$

In the first term we took the derivative with respect to the surface potential, φ_S, since tunneling transmission probability, $\top$, depends sensitively on this potential. In the second term we took the derivative with respect to E_{OL} as the band edge density of states depends on the band alignment. Finally, the subthreshold swing voltage in Eq. (4.7.13) can be expressed in the following form by replacing each term with the appropriate symbol to highlight the four contributing factors:

$$SS = \left(\eta_{el} \times \frac{1}{S_{tunnel}} + \eta_{el} \times \eta_{quant} \times \frac{1}{S_{DOS}} \right)^{-1} \tag{4.7.14a}$$

$$SS = \frac{1}{\eta_{el}} \times \left(\frac{1}{S_{tunnel}} + \frac{\eta_{quant}}{S_{DOS}} \right)^{-1} \tag{4.7.14b}$$

η_{el} and η_{quant} are the electrostatic and quantum confinement efficiencies given by Eq. (4.7.4) and Eq. (4.7.5) respectively. S_{DOS} is the semilog slope of the joint band edge density of states in mV/decade that was discussed in Section 4.3:

$$S_{DOS} \equiv \frac{1}{q} \frac{dE_{OL}}{d\log\left(e^{-|E_{OL}|/qV_0}\right)} = V_0 / \log(e) \tag{4.7.15}$$

Here we redefine S_{tunnel} to be the semilog slope measuring how steeply the tunneling conductance pre-factor changes with respect to the voltage across the body, V_{Body}:

$$S_{tunnel} \equiv \frac{dV_{Body}}{d\log(I_0)} \tag{4.7.16}$$

It is given in mV/decade and I_0 is given by Eq. (4.7.11). S_{tunnel} is the steepness that results from changing the thickness of the tunneling barrier with a changing bias. Since we are interested in the derivative of the log of I_{tunnel}, any value proportional to it can also be used. To good approximation, we can just consider the tunneling probability, $\top$, instead of I_{tunnel} as we did in Section 4.2

A small subthreshold swing voltage can be achieved by having either a small S_{tunnel} or a small S_{DOS} and a good gate efficiency.
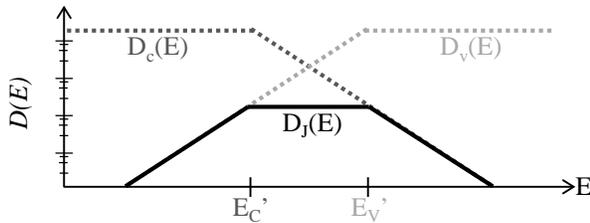
Fig. 4.11: The conduction and valence band density of states, $D_C(E)$ and $D_V(E)$, are shown. Below the band edges the density of states falls off exponentially. The joint density of states, $D_J(E)$ is also shown.

4.7.3   On-State Conductance

Now that we know how to minimize the voltage we need to maximize the on-state conductance. This is discussed thoroughly in two review articles [1, 2] and Chapter 5. Typically there are two methods that are used to increase the conductance. The first is to increase the tunneling area by using a vertical TFET or a bilayer TFET. In these structures, tunneling occurs over a larger overlap region rather than just at the source-channel junction. The second method is to minimize the tunneling barrier. This can be achieved by either reducing the tunneling barrier height or the tunneling barrier width. Oftentimes heavy doping

is used to shrink the tunneling barrier width. As we saw in Section 4.3 this causes S_{DOS} to increase and ruins the subthreshold swing voltage. Because of this, many experimental results with a high conductivity have a terrible subthreshold swing voltage. Consequently, new methods such as the bilayer are needed to control the tunneling barrier width. The tunneling barrier height can also be reduced by using a smaller band gap material or by using a heterojunction to create a smaller effective band gap for tunneling. However, even the minimum effective band gap is limited by the band edge density of states as we saw from Eq. (4.3.1). If we want a particular on-off ratio, the barrier height in the off-state, $E_{g,eff,}$ must be large enough to suppress the band edge density of states, S_{DOS}, by that on-off ratio.

In Section 4.5 we introduced a new method to increase the conductance. When the tunneling probability is low quantum confinement in the direction of tunneling will increase the conductance.

4.8   Maximizing the gate efficiency

In addition to having a small S_{tunnel} or a small S_{DOS} we need to maximize the gate efficiency, η_{el} and η_{quant} to minimize the voltage. Since this is very geometry dependent we will consider the three TFET structures, lateral, vertical and bilayer separately.
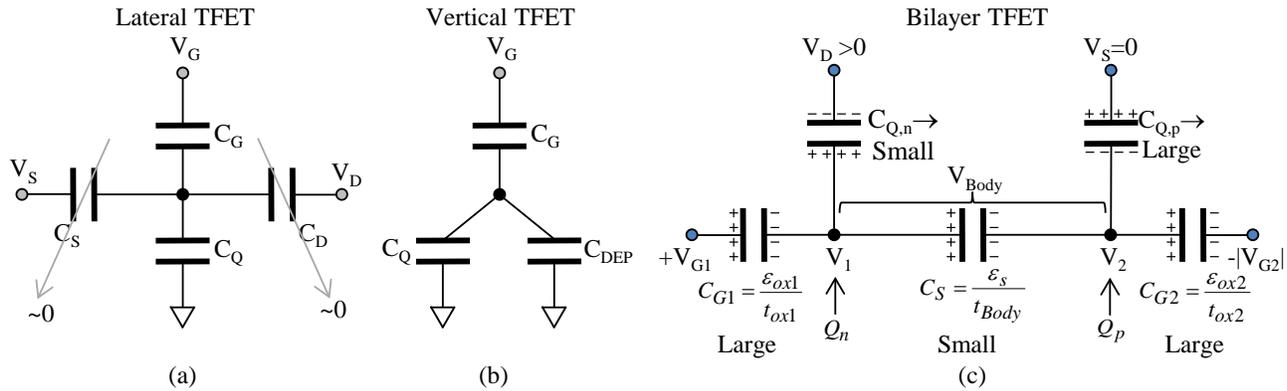
Fig. 4.12: The circuit models for the lateral (a), vertical (b) and bilayer (c) TFETs are shown.

4.8.1   Lateral TFET gate efficiency

In Fig. 4.12(a) we show the circuit model for the lateral TFET. Here we assumed that the body is sufficiently thin that the entire channel will invert. Like a FinFet or nanowire transistor the electrostatics of this device can be very good. If channel is sufficiently long, the gate capacitance, C_G, will be much larger than the source and drain capacitances, C_S and C_D, as only the gate capacitance and the quantum capacitance, C_Q, scale with length. This means that we only need to consider C_G and C_Q to compute the gate efficiency. Since there are two gates, the gate capacitance per unit area is given by:

$$C_G = 2\frac{\varepsilon_{ox}}{t_{ox}} \tag{4.8.1}$$

ε_{ox} is the permittivity of the gate oxide and t_{ox} is the thickness of the gate oxide. The quantum capacitance is simply the voltage needed to add to add more charge to the channel. The charge in the channel, Q_n, is given by the 2D quantum charge and depends on where the Fermi level is relative to the band edge. The charge is given by the following equation for an n-channel device:

$$Q_n = q \times N_{C,2D} \times \ln\left(1 + e^{-\Delta E_{Fn}/k_BT}\right) \text{ where } N_{C,2D} = \frac{m^*_{e,t}}{\pi\hbar^2}k_BT \tag{4.8.2}$$

ΔE_{Fn} is given by $E_C - E_F$. $m^*_{e,t}$ is the electron effective mass in the transverse direction. In computing the quantum capacitance, we assumed a small source drain bias such that we can assume a single Fermi level. We make the same assumption for the vertical and bilayer TFETs as well. The quantum capacitance is given by:

$$C_Q = \frac{dQ_n}{d\phi_s} = -\frac{dQ_n}{d\Delta E_{Fn}} = q \times \frac{m^*_{e,t}}{\pi\hbar^2}\frac{1}{1 + e^{\Delta E_{Fn}/k_BT}} \tag{4.8.3}$$

The overall gate efficiency is just a voltage divider between C_G and C_Q and is given by $\eta_{el} = C_G/(C_Q + C_G)$.

In the subthreshold regime $C_Q \rightarrow 0$ and so the gate efficiency will approach 1 as long as $C_G \gg C_S$ and C_D. In the

overdrive regime, the gate efficiency depends on the Fermi level position and the effective mass. Since the channel needs some charge to conduct current, as given by conventional MOSFET electrostatics, the Fermi level position in the on-state should be set by the minimum required charge to get a given channel conductance. A lower effective mass will also decrease the quantum capacitance and thus increase the gate efficiency in the overdrive regime.

For lateral TFETs the quantum confinement efficiency, η_{quant}, is 1. This is because the confinement energy in the channel is just set by the geometry and does not change with bias.

Overall, we see that a well-designed lateral TFET can have a gate efficiency near 1 in the subthreshold regime, similar to well-designed FinFet and nanowire transistors. In the overdrive regime, the gate efficiency is limited by the quantum capacitance and can be improved by reducing the effective mass and using the minimum required charge in the channel.

4.8.2 Vertical TFET gate efficiency

The key tunneling junction in a vertical TFET is given by the band diagram in Fig. 4.10(h). It can be modeled as a simple MOS capacitor. In doing so, we are neglecting the details of the 2d electrostatics and focusing on only the essential switching action. A fair bit of engineering is required to ensure that vertical TFET actually behaves like a 1d MOS capacitor[41, 42]. If the 2d electrostatics are designed incorrectly different regions of the device can turn on at different biases and smear out the subthreshold swing voltage. Nevertheless, before we even get to that stage of design, we need to understand what the inherent tradeoffs in a vertical architecture are. Consequently, we consider the simplest 1d model shown in Fig. 4.10(h). The circuit model is given in Fig. 4.12(b).

First, we find the quantum confinement efficiency, η_{quant} by finding $dE_{OL}/d\varphi_S$. We choose to measure the potential from the bulk conduction band edge such that the total band bending = φ_S. The overlap energy E_{OL} is given by:

$$E_{OL} = \phi_s - E_G - E_{1e}(\phi_s) \tag{4.8.4}$$

Here E_{1e} is the confinement energy in the triangular well and E_G is the band gap. Plugging this into the definition of η_{quant}, Eq. (4.7.5) we get:

$$\eta_{quant} = \frac{1}{q}\frac{dE_{OL}}{d\phi_s} = 1 - \frac{1}{q}\frac{dE_{1e}}{d\phi_s} \tag{4.8.5}$$

Thus we need to find $dE_{1e}/d\varphi_S$. We can approximate the potential well as a triangular quantum well whose slope is set by the peak electric field in the MOS capacitor. Assuming an infinite triangular well will result in an over-estimate of the confinement energy, but it is sufficient for a first approximation. Consequently, the ground state energy is given by:

$$E_{1e} \approx \left(\frac{9\pi}{8}\right)^{2/3} \times \left(\frac{F^2\hbar^2 q^2}{2m_{e,z}^*}\right)^{1/3} \tag{4.8.6}$$

$m_{e,z}^*$ is the electron effective mass in the tunneling direction. The peak electric field is set by the level of band bending, φ_S, and by the doping, N_d:

$$F = \sqrt{\frac{2qN_d\phi_s}{\varepsilon_s}} \tag{4.8.7}$$

The permittivity of the semiconductor is ε_S. Plugging Eq. (4.8.7) into Eq. (4.8.6) and evaluating $dE_{1e}/d(q \times \varphi_S)$ gives:

$$\frac{dE_{1e}}{d(q \times \phi_s)} = 1 - \eta_{quant} \approx \frac{1}{3}\left(\frac{9\pi}{8}\right)^{2/3} \times \left(\frac{N_d\hbar^2}{m_{e,z}^*\varepsilon_s}\right)^{1/3} \times \phi_s^{-2/3} \tag{4.8.8}$$

Finally, we can use Eq. (4.8.8) in Eq. (4.8.5) to evaluate the quantum efficiency. To estimate η_{quant} we consider Silicon ($m_{e,z}^* = 0.92$) with $E_{OL}=0$ and $N_D=10^{17}$, 10^{18}, 10^{19} and 10^{20}/cm^3. We find $\eta_{quant} = 0.98$, 0.97, 0.93 and 0.88 respectively. In InAs ($m_{e,z}^* = 0.023$), $\eta_{quant} = 0.91$, 0.84 and 0.77 for $N_D=10^{17}$, 10^{18} and 10^{19}/cm^3 respectively. As we can see, lowering the doping will help increase the quantum efficiency, but it will result in a longer depletion width and thus a thicker tunneling barrier and lower current.

Now we can find the electrostatic gate efficiency. The key difference from a lateral TFET is that we have a depletion capacitance in parallel with the quantum capacitance as shown in in Fig. 4.12(b). Once again, η_{el} is given by a

simple voltage divider: $\eta_{el} = C_G / (C_Q + C_{DEP} + C_G)$. The gate capacitance is simply given by $C_G = \varepsilon_{ox}/t_{ox}$. The depletion capacitance is given by:

$$C_{DEP} = \frac{\varepsilon_s}{W_{DEP}} = \sqrt{\frac{q\varepsilon_s N_d}{2\phi_s}} \tag{4.8.9}$$

W_{DEP} is the depletion region width. The quantum capacitance is given by Eq. (4.8.3) but reduced by η_{quant} as the confinement energy level shifts with bias:

$$C_Q = \frac{dQ_n}{d\phi_s} = \frac{-dQ_n}{d\Delta E_{Fn}} \times \frac{d\Delta E_{Fn}}{dE_{OL}} \times \frac{dE_{OL}}{d\phi_s} = q \frac{m^*_{e,t}}{\pi\hbar^2} \frac{1}{1+e^{\Delta E_{Fn}/k_B T}} \times 1 \times \eta_{quant} \tag{4.8.10}$$

Thus we see that in the subthreshold regime the electrostatic efficiency is limited by the depletion capacitance, while in the overdrive regime the quantum capacitance will typically limit the efficiency. In addition to improving η_{quant}, minimizing the doping will reduce the depletion capacitance and thus increase η_{el}. In the subthreshold regime, η_{el} is typically 80-90%. As with the lateral TFET, minimizing the transverse effective mass and the channel charge will result in a lower quantum capacitance and thus a higher overdrive gate efficiency.

Overall, we see that like a conventional planar MOSFET the vertical TFET has subthreshold gate efficiency less than 1 due to the depletion capacitance. Vertical TFETs also suffer slightly from a lower quantum efficiency due to the need for heavy doping to maintain a thin tunneling barrier and a large band bending of at least E_G. Conversely, vertical TFETs do enable a larger tunneling area and thus a higher conductance.

4.8.3   Bilayer TFET gate efficiency

The band diagram for the bilayer TFET is shown in Fig. 4.10(i) and the circuit model is shown in Fig. 4.12(c). As seen in the band diagram the electrons and holes are both quantized, which will result in a lower quantum gate efficiency, η_{quant}, than vertical TFET. Furthermore, since a different voltage is applied to the top and bottom gates, the applied voltage will be split over two gate oxides resulting in a lower electrostatic gate efficiency. Nevertheless, a bilayer structure will allow for the highest on state conductance. Furthermore, as we will see in Section 4.4, not having doping in the tunneling junction will significantly improve the subthreshold swing voltage and should more than compensate for the gate efficiency.

To compute the quantum and electrostatic gate efficiencies we first need to redefine the efficiencies Eq. (4.7.4) and Eq. (4.7.5) to refer to the voltage across the body, V_{body}, rather than the surface potential:

$$\eta_{quant} = \frac{1}{q} \frac{dE_{OL}}{dV_{body}} \tag{4.8.11}$$

$$\eta_{el} = dV_{body} / dV_{G1} \tag{4.8.12}$$

In the lateral and vertical TFETs the potential of the p-side of the junction is fixed by the source and so we only need to know how the surface potential changes. However, in a bilayer TFET, the potential on the p-side, V_2, is not fixed and so it is more convenient to compute the efficiency relative to V_{body}. We also consider the situation where the bias on the n-gate, V_{G1}, is changed while the bias on the p-gate, V_{G2}, is held constant.

We can find η_{quant} by using the definition of E_{OL} to take the derivative of the overlap energy:

$$E_{OL} = qV_{Body} - (E_G + E_{1e} + E_{1h}) \tag{4.8.13}$$

This definition can be seen from Fig. 4.10(i). The triangular well confinement energies are given by:

$$E_{1e} \approx \left(\frac{9\pi}{8}\right)^{2/3} \times \left(\frac{(qV_{Body}/t_{Body})^2 \hbar^2}{2m^*_{e,z}}\right)^{1/3} \quad \text{and} \quad E_{1h} \approx \left(\frac{9\pi}{8}\right)^{2/3} \times \left(\frac{(qV_{Body}/t_{Body})^2 \hbar^2}{2m^*_{h,z}}\right)^{1/3} \tag{4.8.14}$$

Evaluating Eq. (4.8.11) using Eq. (4.8.13) gives:

$$\eta_{quant} = 1 - \frac{2}{3}\left(\frac{9\pi}{8}\right)^{2/3} \times \left(\frac{\hbar^2}{2qt^2_{Body}}\right)^{1/3} \times \left(\frac{1}{\left(m^*_{e,z}\right)^{1/3}} + \frac{1}{\left(m^*_{h,z}\right)^{1/3}}\right) \times \left(V_{Body}\right)^{-1/3} \tag{4.8.15}$$

Next we consider the electrostatic efficiency, η_{el}. It can be optimized by engineering the bilayer thickness and quantum capacitances. Maximizing the electrostatic efficiency means maximizing dV_{Body}/dV_{G1}. As $V_{Body} = V_1 - V_2$, we want to maximize dV_1/dV_{G1} while minimizing dV_2/dV_{G1}. The voltages are labeled on the circuit diagram in Fig. 4.12(c). V_1 and V_2

are the n-channel and p-channel potential, respectively. Consequently, we want the body to be as thick as possible to isolate V_2 from V_{G1} and minimize the body capacitance C_S in the voltage divider. Additionally, we want to minimize the influence of the drain voltage on the n-channel by minimizing the electron quantum capacitance, C_{Qn}, and thus the electron density. Since we want to fix V_2, we want to maximize the influence of the source on the p-channel by maximizing the hole quantum capacitance, C_{Qp}, and increase the hole density until the p-side is degenerate. Therefore, choosing the biases / work functions to control the carrier densities will allow us to improve the electrostatic efficiency, especially when thicker gate oxides are used.

Unfortunately, the carrier density and body thickness are constrained by the required on-state conductance. The fewer electrons present in the channel, the lower the channel conductance and the thicker the body, the lower the tunneling probability. Thus, we need to optimize these tradeoffs to maximize the device performance. Furthermore, computing the electrostatic efficiency, η_{el}, needs to be done numerically as the carrier densities depend on the potentials V_1 and V_2 which both change with gate bias. This is done in detail in [43].

As seen in [43], when the bilayer body thickness is optimized for a reasonable on-state current, Si, Ge and InAs have an overall gate efficiency of 40-50%, with both quantum and electrostatic efficiencies of around 60-70%. While the bilayer has a lower gate efficiency than lateral and vertical TFETs, it will have the highest on-state conductance and it has an undoped tunneling junction which will lead to significantly sharper band edges.

4.9   Other Design Issues to Avoid

When designing a TFET there are several additional issues that can prevent a small subthreshold swing voltage. The first issue that affects many experimental results is trap assisted tunneling. This process occurs when an electron tunnels to a trap in the band gap and is then thermally excited out of the trap. This can result in a temperature dependent subthreshold swing voltage as well as temperature dependent threshold shifts [24, 44-46]. It also increases the subthreshold swing voltage by preventing the tunneling from turning off. High quality interfaces and semiconductors are needed to avoid creating states within the band gap that can lead to trap assisted tunneling.

Another important design issue is to avoid is graded junctions and poor electrostatics. If different regions of the channel start tunneling at different biases, we will get a superposition of I-V curves with different thresholds. This means that the overall subthreshold swing voltage will be smeared out and will be far worse. This can be seen in a variety of simulation studies [41, 42]. Similarly, spatial inhomogeneity can smear out the subthreshold swing voltage.

Finally, a short channel length will result in source to drain tunneling or contact broadening which will increase the subthreshold swing voltage [47]. In order to suppress direct source to drain tunneling, TFETs will need a channel length longer than a corresponding MOSFET in the same material system. This is because TFETs are designed to have a subthreshold swing voltage less than 60 mV/decade and consequently need a stronger suppression of the direct source to drain tunneling.

4.10   Conclusions

After analyzing the different factors that contribute to the operation of a TFET we find that there are 4 key design issues to consider:
1) Modulating the tunneling barrier thickness does not work. It cannot give a steep subthreshold slope at high current density, unless we have an even steeper density of states
2) We must eliminate doping, spatial inhomogeneity, and preserve material quality to get a steep density of states
3) Quantum confinement in the tunneling direction increases the tunneling conductance and 2d-2d tunneling (bilayer type structures) have the highest conductance.
4) Lateral tunneling structures tend to have the best gate efficiency, while bilayer structures have the worst gate efficiency

While it is clear that a steep band edge density of states is needed, to date, there have been no electronic measurements of a steep band edge density of states. Fortunately, optical as well as the electronic measurements available indicate that eliminating doping in the tunneling junction may allow us to achieve the steep density of states required. Bilayer based structures provide an opportunity to achieve this and a high conductance, but unfortunately suffer from a lower gate efficiency. A double gate lateral structure shown in Fig. 4.3 is an alternative that could eliminate doping and potentially have a higher gate efficiency. Unfortunately such a structure will have a lower conductance as it does not take advantage of quantum confinement or the larger tunneling area of the bilayer. To eliminate other forms of spatial inhomogeneity, atomically precise semiconductors such as monolayer semiconductors might be needed.

Overall, there are still some tradeoffs that need to be engineered, but by respecting the design principles above it

should be possible to make a good TFET with a steep subthreshold swing voltage at high current densities.

Acknowledgment

References

[1]   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.

[2]   A. M. Ionescu and H. Riel, "Tunnel Field-Effect Transistors as Energy-Efficient Electronic Switches," *Nature,* vol. 479, pp. 329-337, Nov 2011.

[3]   S. H. Kim, H. Kam, C. Hu, and T.-J. K. Liu, "Germanium-Source Tunnel Field Effect Transistors with Record High I_{ON}/I_{OFF}," presented at the 2009 Symposium on VLSI Technology, Kyoto, Japan, 2009.

[4]   W. Y. Choi, B. G. Park, J. D. Lee, and T. J. K. Liu, "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, W.-Y. Loh, P. Patel, C. Y. Kang, O. Jungwoo, A. Bowonder, *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, K. Donghyun, S. Raghunathan, and K. Saraswat, "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]   E. O. Kane, "Theory of Tunneling," *Journal of Applied Physics,* vol. 32, pp. 83-91, 1961.

[8]   E. O. Kane, "Zener Tunneling in Semiconductors," *Journal of Physics and Chemistry of Solids,* vol. 12, pp. 181-188, 1959.

[9]   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, p. 084507, 2010.

[10]   L. Lattanzio, L. De Michielis, and A. M. Ionescu, "The Electron-Hole Bilayer Tunnel FET," *Solid-State Electronics,* vol. 74, pp. 85-90, Aug 2012.

[11]   J. T. Teherani, S. Agarwal, E. Yablonovitch, J. L. Hoyt, and D. A. Antoniadis, "Impact of Quantization Energy and Gate Leakage in Bilayer Tunneling Transistors," *IEEE Electron Device Letters,* vol. 34, pp. 298-300, Feb 2013.

[12]   G. Dewey, B. Chu-Kung, J. Boardman, J. M. Fastenau, J. Kavalieros, R. Kotlyar, *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),* p. 33.6, 2011 2011.

[13]   J. Knoch, S. Mantl, and J. Appenzeller, "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.

[14]   J. D. Dow and D. Redfield, "Toward a Unified Theory of Urbach's Rule and Exponential Absorption Edges," *Physical Review B,* vol. 5, p. 594, 1972.

[15]   S. Agarwal and E. Yablonovitch, "Band-Edge Steepness Obtained from Esaki and Backward Diode Current-Voltage Characteristics," *Submitted for Publication,* 2013.

[16]   S. R. Johnson and T. Tiedje, "Temperature Dependence of the Urbach edge in GaAs," *Journal of Applied Physics,* vol. 78, pp. 5609-5613, 1995.

[17]   T. Tiedje, E. Yablonovitch, G. D. Cody, and B. G. Brooks, "Limiting Efficiency of Silicon Solar Cells," *IEEE Transactions on Electron Devices,* vol. 31, pp. 711-716, 1984.

[18]   J. I. Pankove, "Absorption Edge of Impure Gallium Arsenide," *Physical Review,* vol. 140, pp. A2059-A2065, 1965.

[19]   P. Van Mieghem, "Theory of Band Tails in Heavily Doped Semiconductors," *Reviews of Modern Physics,* vol. 64, pp. 755-793, 1992.

[20]   E. O. Kane, "Band Tails in Semiconductors," *Solid-State Electronics,* vol. 28, pp. 3-10, 1985.

[21]   Z. Zhang, R. Rajavel, P. Deelman, and P. Fay, "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.

[22]   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.

[23]   D. Pawlik, B. Romanczyk, P. Thomas, S. Rommel, M. Edirisooriya, R. Contreras-Guerrero, *et al.*, "Benchmarking and Improving III-V Esaki Diode Performance with a Record 2.2 MA/cm^2 Peak Current Density to Enhance TFET Drive Current," in *Electron Devices Meeting (IEDM), 2012 IEEE International*, 2012, pp. 27.1.1-27.1.3.

[24]   S. Mookerjea, D. Mohata, T. Mayer, V. Narayanan, and S. Datta, "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.

[25]   S. Agarwal and E. Yablonovitch. (2011, September). Pronounced Effect of pn-Junction Dimensionality on Tunnel Switch Sharpness. *eprint arXiv:1109.0096*. Available: http://arxiv.org/abs/1109.0096

[26]   K. Tomioka, M. Yoshimura, and T. Fukui, "Steep-Slope Tunnel Field-Effect Transistors Using III-V Nanowire/Si Heterojunction," in *2012 Symposium on VLSI Technology*, 2012, pp. 47-48.

[27]   J. Appenzeller, Y. M. Lin, J. Knoch, and P. Avouris, "Band-to-Band Tunneling in Carbon Nanotube Field-Effect Transistors," *Physical Review Letters,* vol. 93, p. 196805, 2004.

[28]   J. Bardeen, "Tunneling From a Many Particle Point of View," *Physical Review Letters,* vol. 6, pp. 57-59, 1961.

[29]   C. B. Duke, *Tunneling in Solids*. New York: Academic Press, Inc, 1969.

[30]   W. A. Harrison, "Tunneling from an Independent-Particle Point of View," *Physical Review,* vol. 123, pp. 85-89, 1 July 1961.

[31]   J. R. Oppenheimer, "Three Notes on the Quantum Theory of Aperiodic Effects," *Physical Review,* vol. 31, pp. 66-81, Jan 1928.

[32]   A. F. J. Levi, *Applied Quantum Mechanics*: Cambridge University Press, 2006.

[33]   G. Gamow, "Zur Quantentheorie des Atomkernes," *Z. Physik,* vol. 51, p. 204, 1928.

[34]   M. A. Kastner, "The single-electron transistor," *Reviews of Modern Physics,* vol. 64, p. 849, 1992.

[35]   S. Datta, *Quantum Transport : Atom to Transistor*. Cambridge, UK: Cambridge University Press, 2005.

[36]   P. Patel, "Steep Turn On/Off "Green" Tunnel Transistors," PhD, Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, 2010.

[37]   R. Li, Y. Q. Lu, G. L. Zhou, Q. M. Liu, S. D. Chae, T. Vasen*, et al.*, "AlGaSb/InAs Tunnel Field-Effect Transistor With On-Current of 78 uA/um at 0.5 V," *IEEE Electron Device Letters,* vol. 33, pp. 363-365, Mar 2012.

[38]   M. Sweeny and J. M. Xu, "Resonant Interband Tunnel-Diodes," *Applied Physics Letters,* vol. 54, pp. 546-548, Feb 1989.

[39]   S. L. Rommel, T. E. Dillon, M. W. Dashiell, H. Feng, J. Kolodzey, P. R. Berger*, et al.*, "Room Temperature Operation of Epitaxially Grown $Si/Si_{0.5}Ge_{0.5}/Si$ Resonant Interband Tunneling Diodes," *Applied Physics Letters,* vol. 73, pp. 2191-2193, 1998.

[40]   S. Krishnamoorthy, P. S. Park, and S. Rajan, "Demonstration of forward inter-band tunneling in GaN by polarization engineering," *Applied Physics Letters,* vol. 99, pp. 233504-3, 2011.

[41]   S. H. Kim, S. Agarwal, Z. A. Jacobson, P. Matheu, H. Chenming, and L. Tsu-Jae King, "Tunnel Field Effect Transistor With Raised Germanium Source," *Electron Device Letters, IEEE,* vol. 31, pp. 1107-1109, 2010.

[42]   Y. Lu, G. Zhou, R. Li, Q. Liu, Q. Zhang, T. Vasen*, et al.*, "Performance of AlGaSb/InAs TFETs With Gate Electric Field and Tunneling Direction Aligned," *Electron Device Letters, IEEE,* vol. 33, pp. 655-657, 2012.

[43]   S. Agarwal, J. T. teherani, J. L. Hoyt, D. A. Antoniadis, and E. Yablonovitch, "Engineering the Electron-Hole Bilayer Tunneling Field-Effect Transistor," *Submitted for Publication,* 2013.

[44]   G. A. M. Hurkx, D. B. M. Klaassen, and M. P. G. Knuvers, "A New Recombination Model for Device Simulation Including Tunneling," *Electron Devices, IEEE Transactions on,* vol. 39, pp. 331-338, 1992.

[45]   G. A. M. Hurkx, D. B. M. Klaassen, M. P. G. Knuvers, and F. G. O'Hara, "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.

[46]   J. Furlan, "Tunnelling Generation-Recombination Currents in a-Si Junctions," *Progress in Quantum Electronics,* vol. 25, pp. 55-96, 2001.

[47]   K. Ganapathi, Y. Yoon, and S. Salahuddin, "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.