A Quantum Mechanical Analysis of Gate Tunneling Current in MOSFETs

Lawrence Liu
Department of Electrical and Computer Engineering,
University of California–Los Angeles, Los Angeles, California 90095, USA

This paper provides a brief overview of the operation of MOSFETs and their applications in CMOS logic gates. The equations of the approximate operation characteristics of MOSFETs are presented, and then they are analyzed to show how scaling the MOSFET oxide thickness affects the operation of the MOSFET. We then develop a model for the tunneling current through the oxide layer of a MOSFET through Bardeen Tunneling Theory, which we used to show that the tunneling current scales exponentially with decreasing oxide thickness. We then analyze the effects of tunneling current on the operation of a MOSFET, and discuss ways to avoid the effects of tunneling current.

I. INTRODUCTION

Metal Oxide Semiconductor Field Effect Transistors (MOSFETs) are the most common type of transistor used in modern electronics, and at the heart of modern microprocessors. Let us start by briefly discussing the basic operation of a MOSFETs and a common application of MOSFETs in Complemetary Metal Oxide Semiconductor (CMOS) logic gates.

A. Basic MOSFET Structure

The MOSFET is based around a Metal Oxide Semiconductor (MOS) capacitor. Historically this was a gate made out of a metal plate on top of a semiconductor substrate, which was separated by an insulating layer of oxide, traditionally silicon dioxide SO_2.

Now to make a MOSFET we add a source and drain silicon regions to the silicon substrate. The source and drain regions are doped with impurities to be the opposite type to that of the base silicon substrate. For example, if the base silicon substrate is p-type, then the source and drain regions are n-type, and vice versa. We call the MOSFET with n-type source and drain regions an nMOSFET, and the MOSFET with p-type source and drain regions. We have drawn both an nMOSFET and a pMOSFET in Figure (1 and 2) along with their respective circuit symbols.

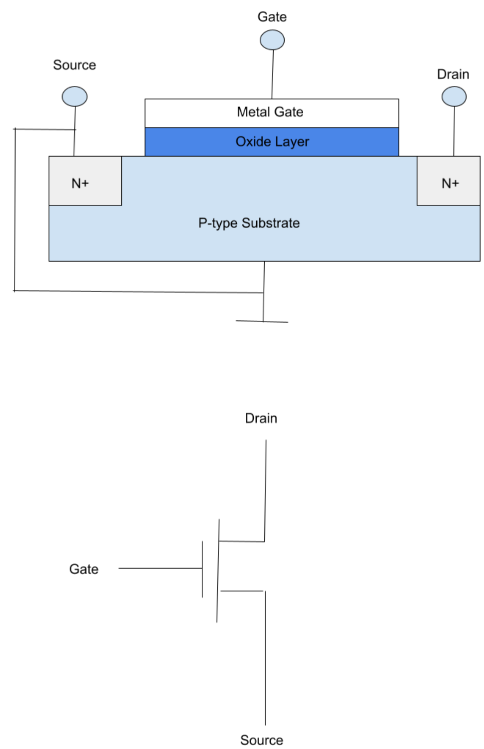

FIG. 1: A diagram of an nMOSFET and its circuit symbol.

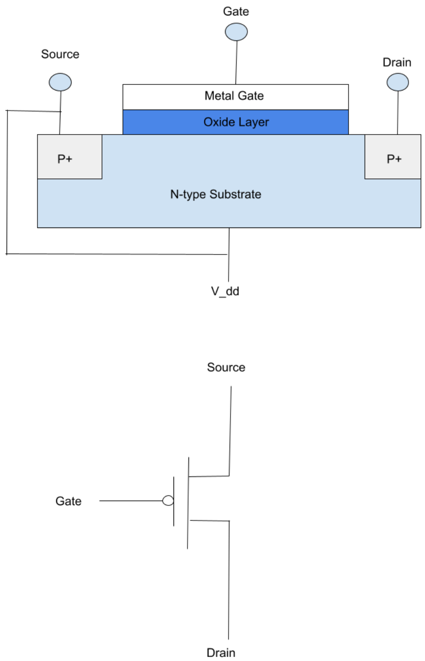

FIG. 2: A diagram of an pMOSFET and its circuit symbol.

Let us define V_{ds} as the voltage between the drain and source, and V_{gs} as the voltage between the gate and source.

B. MOSFET Operation

Now let us consider a the operation of a nMOSFET in a qualitative sense. If we apply a positive voltage to the gate, then the holes in the p-type substrate will be pushed away from the gate, or more accurately, the minority electrons will be drawn to the gate. If enough are drawn to the gate, then an inversion layer of n-type Silicon will form at the surface of the substrate since enough electrons will be present to become the majority carrier. This inversion layer will form a conductive channel between the source and drain, allowing current to flow between the source and drain. Therefore in a very rough sense, we can see that a MOSFET acts as a voltage controlled switch.

In a more rigorous sense, what happens is that the gate voltage "bends" the energy bands in the p-type substrate lower. Once these bands bend to the degree that near the surface of the substrate, the conduction band is closer to the Fermi level than the valence band, then the inversion layer will form. This threshold is given by [2]

$$V_T = V_{FB} + 2\phi_B + \frac{Q_{SS}}{C_{OX}} \tag{1}$$

Where V_{FB} is the flatband voltage, ie the difference in the work functions of the gate and the substrate, ϕ_B is the difference between the Fermi level and the intrinsic Fermi level divided by the electron charge, $Q_{SS} = \sqrt{2\epsilon_s q N_a(2\phi_B)}$ where N_a is the doping concentration of the substrate, and C_{OX} is the capacitance of the oxide layer per unit area.

We can see that the opposite happens for a pMOSFET. If we apply a negative V_{GS}, then since the electrons are pushed away and the holes are drawn to the gate, then an inversion layer of p-type Silicon will form at the surface of the substrate. Or from a bands perspective, the bands will be "bended" upwards to the degree that the valence band is closer to the Fermi level than the conduction band. We have that this threshold is given by [2]

$$V_T = V_{FB} - 2\phi_B - \frac{2\epsilon_s q N_d(2\phi_B)}{C_{OX}} \tag{2}$$

Where N_d is the doping concentration of the substrate.

Thus when we a apply a voltage greater or less than the threshold voltage for the nMOSFET and pMOSFET respectively, then the inversion layer will form. We can view this layer as a effectively a resistor between the source and drain. However as we increase the current across the drain and the source V_{DS}, we will start to experience "pinch off" where the inversion layer will start to narrow. Once the voltage is high enough, the inversion layer will pinch off completely, and no longer connect the source and drain. This is depicted in Figure 3. This will cause the MOSFET to no longer act as a voltage controlled resistor between the source and drain, and instead act as a voltage controlled current source. We call this operation region the "saturation" region as opposed the "Ohmic" region where the MOSFET acts as a voltage controlled resistor.

The threshold for V_{DS} at which the MOSFET enters the saturation region can be approximated by [3]

$$V_{DS,sat} = V_{GS} - V_T \tag{3}$$

The current in the saturation region therefore can be approximated by[3]

$$I_{D,sat} = \frac{1}{2} C_{OX} \frac{W}{L} (V_{GS} - V_T)^2 \tag{4}$$

Where W is the width of the MOSFET and L is the length of the MOSFET.

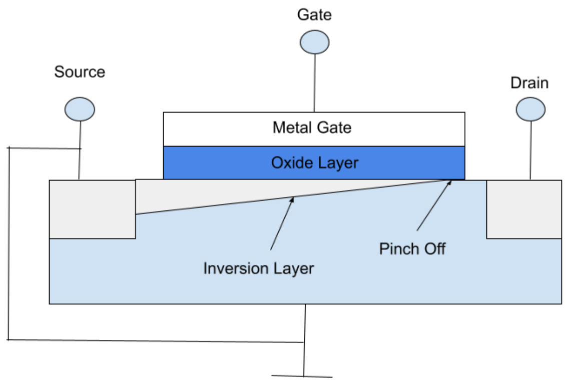

FIG. 3: Pinch off in a MOSFET.

C. Examples of MOSFET based circuits

Now with this understanding of the operation of a MOSFET, we can now look at some examples of MOSFET based circuits.

1. *CMOS Inverter*

By connecting a pMOSFET and a nMOSFET in series, we can create a CMOS inverter. This is depicted in Figure 4.

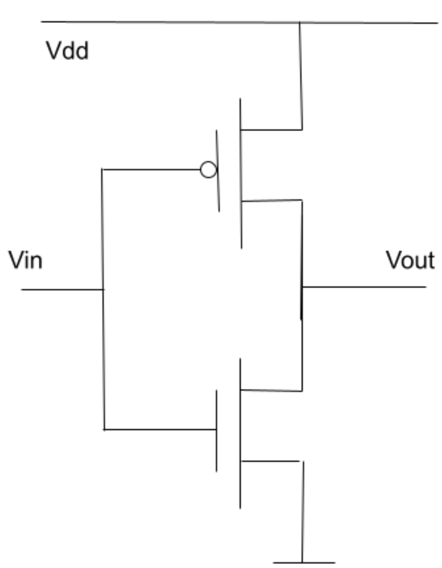

FIG. 4: CMOS inverter.

To understand this circuit let us consider two cases when the input is "low" and when the input is "high". Let us assume that that the threshold voltage of the nMOSFET is equal to the threshold voltage of the pMOSFET, then in the case where the input is low, then the nMOSFET will be effectively an open circuit, and the pMOSFET will be effectively a closed circuit. Therefore the output will be high. Conversely, if the input is high, then the nMOSFET will be effectively a closed circuit, and the pMOSFET will be effectively an open circuit. Therefore the output will be low. Thus we can see that this circuit acts as an inverter. Likewise we can create more complex cirucits to implement logic gates such as a NAND or NOR gate using CMOS, which are discussed in Appendix A.

D. Speed

We should note that a MOSFET is not an instantaneous switch, we can create an basic approximation by considering the MOSFET as a RC circuit. The capacitance comes from the gate capacitance of the next inverter along with the parasitic capacitance coming from the various wires connecting the MOSFETs, and the capacitance that is formed in the PN junction between the source and drain and the substrate in the driving MOSFETs. We can add this capacitance to the MOSFET model as depicted in figure (5) for a CMOS inverter train.

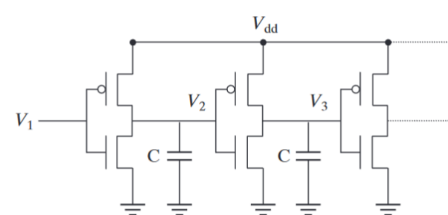

FIG. 5: CMOS inverter train, figure from *Modern Semiconductor Devices for Integrated Circuits* by Chenming C. Hu. [3]

We can approximate the switching resistance of the MOSFET as $R \approx \frac{V_{dd}}{2I_{on}}$, Thus we have that the switching delay is proportional to the RC time constant of the MOSFET. We have that:

$$\tau \propto \frac{C}{I_{on}} V_{DD} \tag{5}$$

II. BENEFITS OF SHRINKING GATE OXIDE LAYER

Now armed with the knowledge of how MOSFETs work, let us examine how a thinner gate oxide layer impacts the performance of a MOSFET. We have that the capacitance per unit area of the oxide layer is given by

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} \tag{6}$$

Where t_{ox} is the thickness of the oxide layer and the permittivity of the oxide is ϵ_{ox}. Now if we consider a MOSFET with a thinner oxide layer αt_{ox} we have that the capacitance of the oxide layer is given by

$$C_{ox} = \frac{\epsilon_{ox}}{\alpha t_{ox}} \tag{7}$$

Therefore we can see that the unit capacitance of the oxide layer is inversely proportional to the thickness of the oxide layer. Now if we return to equations (1) and (2), we can see that the threshold voltage is varies inversely proportional to the unit capacitance of the oxide layer. Thus a thinner oxide layer will have a lower threshold voltage. This is important because a lower threshold voltage means that the MOSFET will require less voltage, and therefore at a same current, will dissipate less power.[3]

This will also affect the speed of the MOSFET by increasing the on current. We have that the on current given by equation (4) is proportional to $(V_{GS} - V_T)^2$. Therefore a lower threshold voltage will result in a higher on current. Now if we recall that the switch time of a MOSFET is proportional to $\frac{1}{I_{on}}$. Therefore a smaller threshold voltage will result in a faster switch time, or alternatively a lower power dissipation for the same switching time.[3]

A. Issues with MOSFET Scaling

A question that naturally arises is what would stop us from making our Gate Oxide Layer thinner and thinner? Well, there will be several issues that will arise as we make our MOSFETs smaller. The most obvious issue is that it will be harder to manufacture thinner and thinner gate oxide layers. Moreover, there are also physical limitations that will arise as we make our MOSFETs smaller. With large size MOSFETs, we can treat the MOS capacitor as a classical capacitor, however, as we make our MOSFETs smaller, we will have to consider quantum mechanical effects. In particular, we will have to consider Quantum Tunneling. The next section will explain how Quantum Tunneling will affect the behavior of the MOSFET, and discuss how we can mitigate this effect.

III. BARDEEN TUNNELING THEORY

Let us start by briefly summarizing Bardeen Tunneling Theory. Let us consider a quantum mechanical system as depicted in figure (6) consisting of two potential wells separated by a barrier:

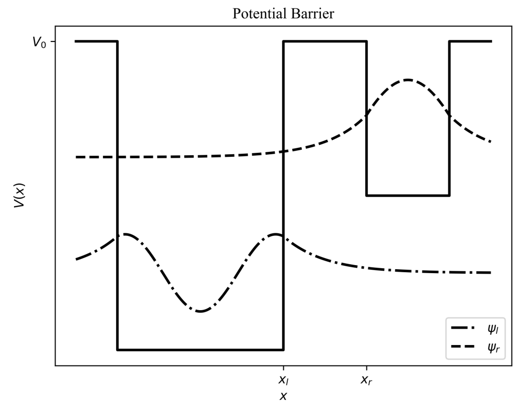

FIG. 6: Potential wells and wavefunctions.

As we depicted, let us assume that both of these potential wells are symmetric around their own origins, and thus the potential outside the well is the same, which we defined in figure (6) to be V_0. Then we have that the potential is given by

$$V(x) = \begin{cases} V_l(x) & x < x_1 \\ V_r(x) & x \geq x_1 \end{cases} \tag{8}$$

Where $V_l(x)$ is the potential function for the left well, and $V_r(x)$ is the potential function for the right well, and x_1 is some arbitrary point $x_l < x_1 < x_r$. We can see that we can express this as:

$$V(x) = V_l(x) - \Theta(x - x_1)(V_0 - V_r(x)) \tag{9}$$

Where $\Theta(x)$ is the Heaviside step function. Therefore we can see that we can write the hamiltonian for this system as:

$$H = \frac{p^2}{2m} + V_l(x) - \Theta(x - x_1)(V_0 - V_r(x)) \tag{10}$$

We can see that if we define H_l as the hamiltonian of a system consisting solely of the left well, and $H' = \theta(x - x_1)(V_0 - V_r(x))$, then we can write the hamiltonian as:

$$H = H_l + H' \tag{11}$$

Therefore we can see the that the tunneling rate can be modeled by with Fermi's Golden Rule, with H' as the perturbation that turns on at $t = 0$. If we make the key assumption that the energy of the left well is approximately equal to the energy of the right well, ie $E_l \approx E_r = E$. We have that the transition rate is given by:

$$\Gamma_{l \to r} = \frac{2\pi}{\hbar} |\langle r| H' |l\rangle|^2 \rho(E) \tag{12}$$

We have that the matrix element is given by:

$$\langle r|\, H' \,|l\rangle = \int_{-\infty}^{\infty} \psi_r^*(x) H' \psi_l(x) dx \tag{13}$$

After doing some simplifications derived in Appendix B, and with the assumption that $E_l \approx E_r = E$ we have that the matrix element is given by:

$$\langle r|\, H' \,|l\rangle = \frac{\hbar^2}{2m}\left(\psi_r^* \frac{d}{dx}\psi_l - \psi_l \frac{d}{dx}\psi_r^*\right)\Bigg|_{x=x_1} \tag{14}$$

Which is the central result of Bardeen Tunneling Theory. [5].

A. Application to MOSFETs

Now let us use it to try to approximate the gate source tunneling current of a MOSFET. We model the potentials and for a electron in the gate and substrate as two finite square wells, as depicted in figure (6), and with the raised potential barrier as a crude model of the oxide layer. Recall that wavefunction for a particle in a finite square well of length L centered at $x = 0$ is given by:

$$\psi(x) = \begin{cases} A\frac{e^{-K\frac{L}{2}}}{\cos\left(K\frac{L}{2}\right)}\cos(Kx) & |x| < \frac{L}{2} \\ Ae^{-k|x|} & |x| > \frac{L}{2} \end{cases} \tag{15}$$

Where $K = \frac{1}{\hbar}\sqrt{2m(V-E)}$, $k = \frac{1}{\hbar}\sqrt{2mE}$, and A is a normalization constant, and V is the height of the walls of the box. Then we have that the matrix element is given by:

$$\langle r|\, H' \,|l\rangle = \frac{\hbar^2}{2m}\Bigg(A_r e^{-k_r x}\frac{d}{dx}A_l e^{k_l\left(x-L_b-\frac{L_l+L_r}{2}\right)} - A_l e^{k_l\left(x-L_b-\frac{L_l+L_r}{2}\right)}\frac{d}{dx}A_r e^{-k_r x}\Bigg) \tag{16}$$

Where L_b is the length of the barrier, L_l is the length of the left well, and L_r is the length of the right well. Likewise if we note that since we assumed with the Bardeen Tunneling Theory that the energy of the particle in the left well is approximately equal to the energy of the particle in the right well, we have that $k_l \approx k_r = k$. Thus we have that the matrix element is given by:

$$\langle r|\, H' \,|l\rangle = -\frac{\hbar^2}{m}A_l A_r k e^{-k\left(L_b+\frac{L_l+L_r}{2}\right)} \tag{17}$$

Therefore we can see that the tunneling rate is given by:

$$\Gamma_{l\to r} = \frac{2\pi}{\hbar}\frac{\hbar^4}{m^2}A_l^2 A_r^2 k^2 e^{-2k\left(L_b+\frac{L_l+L_r}{2}\right)}\rho(E) \tag{18}$$

Now for our case, since the barrier is the dielectric layer, we have that $L_b = t_{ox}$. Furthermore, we have that the tunneling current is given by:

$$I_{tunnel} = q\Gamma_{l\to r} \tag{19}$$

Thus we can see that from this crude model of the MOSFET, we can see that the tunneling current scales exponentially with the $-t_{ox}$. This is reflected in the actual data of the gate source tunneling current compared with the oxide thickness, as we showed in figure (7). This approach for modeling the tunneling current was based on the paper *Theory of direct tunneling current in metal-oxide-semiconductor structures* by Clerc et al. [1]

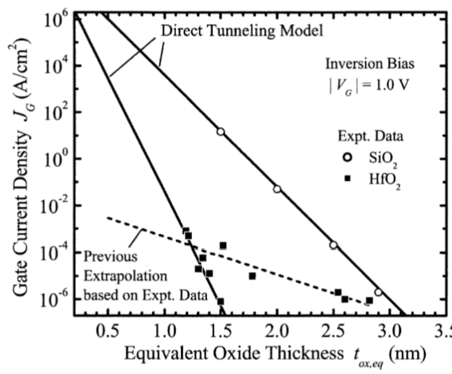

FIG. 7: Gate Source Tunneling Current Density vs Oxide Thickness, figure from *MOSFET Gate Leakage Modeling and Selection Guide for Alternative Gate Dielectrics Based on Leakage Considerations* by Yee-Chia Yeo, Tsu-Jae King, and Chenming Hu [7]

B. Consequences of Gate Source Tunneling and Mitigation Methods

As was shown in figure (7), at 1.2nm the gate source leakage current is roughly $10^3 A/cm^2$. Thus if we had a chip that consisted of MOSFETs with a total area of $1mm^2$, we would have a leakage current of $10A$. This would drain a phone battery (which has a capacity of $3000mAh$) in 10 minutes. Therefore it is not suitable to continue to scale down the oxide thickness. Thus we want to find ways to mitigate the gate source tunneling current, while still keeping the benefits of the resulting higher C_{ox} we discussed earlier.

A common solution to this problem is to use a high-k dielectric,[4] such as HfO_2 instead of the traditional SiO_2. HfO_2 has a dielectric constant that is roughly 6 times that of SiO_2, thus we can see that we can get the same C_{ox} with a 6 times thicker HfO_2 layer. However since the tunneling current scales exponentially with the thickness of the oxide layer, this will have a significantly smaller gate source leakage, which can also be seen in figure (7).

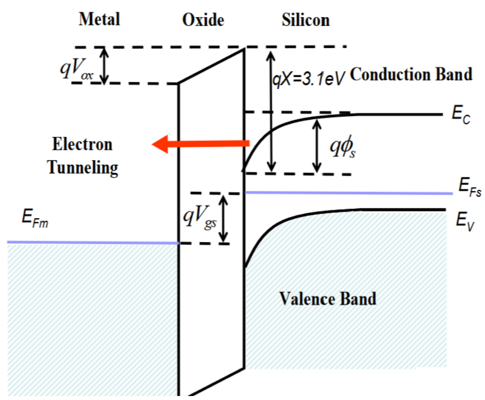

FIG. 8: Trapezoidal model for the Gate potential, figure from *Quantum Mechanical Effects on Mosfet Scaling Limit* by Lihui Wan [6]

C. Ways to Improve the Model

While our model was able to capture the exponential dependence of the gate source tunneling current on the oxide thickness, it is not able to capture the magnitude of the gate source tunneling current. This is because we made a number of simplifying factors, most importantly we assumed the potential in the gate oxide is flat, however we are applying a voltage across the gate oxide, so the barrier will be trapezoidal and not rectangular as depicted in figure (8). To model these effects we would need to use WKB approximation, which was beyond the scope of this class and this paper, however the predictions of such a model is depicted as the Direct Tunneling Model line in figure (7).

Likewise we also only considered direct tunneling, however at higher thicknesses other effects begin to dominate such as trap-assisted tunneling. This is show in the figure (7), where we can see that for high thicknesses of HfO_2, the gate current density no longer begins to behave according the direct tunneling models.[7]

IV. CONCLUSION

In this paper we have discussed the physics of MOSFETs, and how they are used in modern day electronics. We then discussed the performance benefits of a thinner oxide layer, as well as the drawbacks of a thinner oxide layer, namely gate source tunneling. We then discussed the physics of gate source tunneling, derived the Bardeen Tunneling Theory, and a model for the gate source tunneling current based on the Bardeen Tunneling Theory. Finally we discussed the consequences of this tunneling current, and how to mitigate it.

[1] R. Clerc, A. Spinelli, G. Ghibaudo, and G. Pananakakis. Theory of direct tunneling current in metal-oxide-semiconductor structures. *Journal of Applied Physics*, 91(3):1400–1409, 02 2002.

[2] C. Hu. *Chapter 5: MOS Capacitor*. Prentice Hall, 2010.

[3] C. Hu. *Chapter 6: MOS Transistor*. Prentice Hall, 2010.

[4] C. Hu. *Chapter 7: MOSFETs in ICs—Scaling, Leakage, and Other Topics*. Prentice Hall, 2010.

[5] H. J. Reittu. Fermi's golden rule and Bardeen's tunneling theory. *American Journal of Physics*, 63(10):940–944, 10 1995.

[6] L. Wang. *Quantum Mechanical Effects on Mosfet Scaling Limit*. PhD thesis, Georgia Institute of Technology, 2006.

[7] Y.-C. Yeo, T.-J. King, and C. Hu. Mosfet gate leakage modeling and selection guide for alternative gate dielectrics based on leakage considerations. *IEEE Transactions on Electron Devices*, 50(4):1027–1035, 2003.

Appendix A: MOSFET Logic Gates

1. NAND Gate

Here we discuss an example of a CMOS logic gate, the NAND gate. The NAND gate is a universal gate, meaning that any boolean function can be constructed using only NAND gates. We have drawn an example of a NAND gate in figure (9).

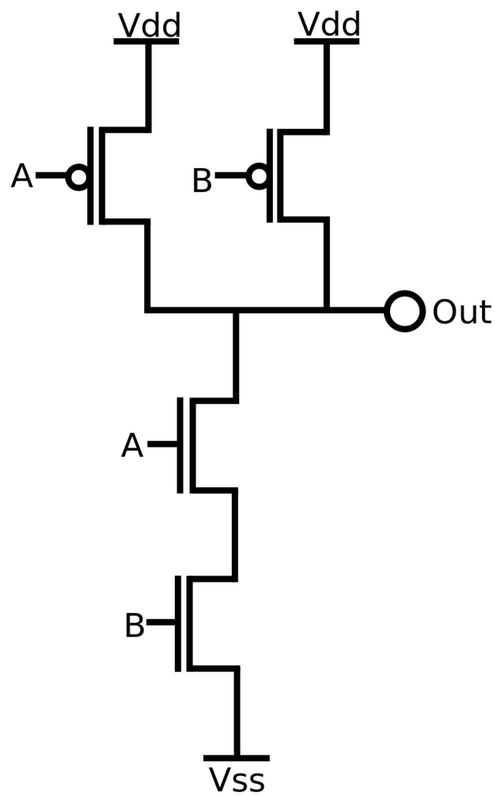

FIG. 9: NAND Gate, figure By Justin-Force - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=2593317

As we can see, the NAND gate consists of two PMOS transistors in parallel with two NMOS transistors in series. Thus if both A and B are high, then both NMOS transistors will be on, and the PMOS transistors will be off, thus the output will be low. If either A or B or both are low, then at least one of the NMOS transistors will be off and at least one of the PMOS transistors will be on, thus the output will be high.

Appendix B: Derivation of the Bardeen Tunneling Theory

We have that the matrix element as we established previously is

$$\langle r| H' |l\rangle = \int_{-\infty}^{\infty} \psi_r^*(x) H' \psi_l(x) dx \tag{B1}$$

However since $H' = -\Theta(x-x_1)(V_0 - V_r(x))$ we have that

$$\langle r| H' |l\rangle = \int_{x_1}^{\infty} \psi_r^*(x) H' \psi_l(x) dx \tag{B2}$$

We also have that $H' = H_l - H$ therefore we get:

$$\langle r| H' |l\rangle = \int_{x_1}^{\infty} \psi_r^*(x)(H_l - H)\psi_l(x) dx \tag{B3}$$

$$= \int_{x_1}^{\infty} \psi_r^*(x)(H_l + \frac{\hbar^2}{2m}\frac{d^2}{dx^2} - V_r(x))\psi_l(x) dx \tag{B4}$$

$$= \int_{x_1}^{\infty} \psi_r^*(x)(E_l + \frac{\hbar^2}{2m}\frac{d^2}{dx^2} - V_r(x))\psi_l(x) dx \tag{B5}$$

If we perform integration by parts we get:

$$\begin{aligned}\int \psi_r^*(x)\frac{d^2}{dx^2}\psi_l(x)dx =& \psi_r^*(x)\frac{d}{dx}\psi_l(x) - \int \frac{d}{dx}\psi_r^*(x)\frac{d}{dx}\psi_l(x)dx \\ =& \psi_r^*(x)\frac{d}{dx}\psi_l(x) - \psi_l(x)\frac{d}{dx}\psi_r^*(x) \\ &+ \int \psi_l(x)\frac{d^2}{dx^2}\psi_r^*(x)dx\end{aligned}$$

Therefore since $\frac{d}{dx}\psi_l\big|_{x=\infty} = \frac{d}{dx}\psi_r\big|_{x=\infty} = 0$ we get:

$$\begin{aligned}\langle r| H' |l\rangle =& \frac{\hbar^2}{2m}\left(\psi_r^*(x)\frac{d}{dx}\psi_l(x) - \psi_l(x)\frac{d}{dx}\psi_r^*(x)\right)\Big|_{x=x_1} \\ &- E_l\int_{x_1}^{\infty}\psi_r^*(x)\psi_l(x)dx + \int_{x_1}^{\infty}\psi_l(x)H_r\psi_r^*(x)dx \\ =& \frac{\hbar^2}{2m}\left(\psi_r^*(x)\frac{d}{dx}\psi_l(x) - \psi_l(x)\frac{d}{dx}\psi_r^*(x)\right)\Big|_{x=x_1} \\ &+ (E_r - E_l)\int_{x_1}^{\infty}\psi_r^*(x)\psi_l(x)dx\end{aligned}$$

Since we assume that $E_r = E_l$ we get:

$$\langle r| H' |l\rangle = \frac{\hbar^2}{2m}\left(\psi_r^*(x)\frac{d}{dx}\psi_l(x) - \psi_l(x)\frac{d}{dx}\psi_r^*(x)\right)\Big|_{x=x_1} \tag{B6}$$