University of Seville
Seville, June 2019

A study on fundamental physical limitations of CMOS technologies

Rita González Márquez

Supervisors: Jorge Fernández Berni, Rocío del Río Fernández

Bachelor Thesis
Degree in Physics

Department of Electronics and Electromagnetism
Faculty of Physics
UNIVERSITY OF SEVILLE

Acknowledgements

First of all, I would like to thank my family for being a, really needed, support throughout the process of making this thesis and the four last years. Also, to my friends, who brightened the worst moments and made out of the most stressful days memorable ones.

I would also like to express my gratitude towards my supervisors: Jorge Fernández Berni and Rocío del Río Fernández, for the time and the effort dedicated to my work.

I want to thank Rocío for her wise advices, for her support and for her encouragement to the realization of this bachelor thesis. Without her participation and collaboration, this work would not have been what it is today.

I want to thank Jorge for the interest he has always shown, the endless hours dedicated to my work, for listening -not hearing- everything that I wanted to say and discussing with me with infinite patience. You have taught me to seek for the true answers no matter what and the importance of honesty and moral in research, a quality not always present these days.

Thanks to both of them, I was allowed to deeply enjoy the process of making this thesis and the research behind it, and I will be forever grateful for that.

University of Seville

Sevilla, July 2019

Abstract

In this work the problem of power dissipation for nano-CMOS technologies is addressed. The main focus is to analyze the physics behind the impossibility of scaling the subthreshold slope below 60 mV/decade. For that an NMOS transistor is studied. First, the carrier transport processes in subthreshold region are analyzed to understand the thermionic nature of their emission over the channel potential barrier. Then, the drain-source current is derived step by step from a diffusion current expression, procedure that has not been reported in the literature up to now. With that, it was proven that the subthreshold current is in fact a diffusion current and that the amount of current through the channel depends on the fraction of carriers able to thermionically surpass the barrier. From the subthreshold drain-source current, an analytical expression for the subthreshold slope SS is obtained, making use of its definition as the inverse of the transfer characteristic's slope. After that, the different terms of SS are analyzed separately, and possible approaches to reduce its value are discussed. The fundamental nature of the 60 mV/dec limitation is proved to be a thermionic emission limitation. Finally, some emerging devices and their subthreshold slopes are analyzed and compared to the traditional MOSFET.

Keywords – Integrated Circuits, CMOS, Power Dissipation, Subthreshold Slope, Thermionic Emission, Diffusion Current

Contents

List of Figures

List of Tables

List of Acronyms

- MOS: Metal-Oxide Semiconductor
- FET: Field-Effect Transistor
- MOSFET: Metal-Oxide Semiconductor Field-Effect Transistor
- CMOS: Complementary Metal-Oxide Semiconductor
- NMOS: N-type Metal-Oxide Semiconductor
- PMOS: P-type Metal-Oxide Semiconductor
- SS: Subthreshold Slope / Subthreshold Swing
- IC: Integrated Circuit
- FO: Fan-Out
- BJT: Bipolar Junction Transistor
- SOI: Silicon on Insulator
- FinFET: Fin Field-Effect Transistor
- TFET: Tunneling Field-Effect Transistor / Tunnel Field-Effect Transistor
- ISSCC: International Solid-State Circuits Conference
- MPU: Microprocessor Unit
- DRAM: Dynamic Random Access Memory
- DSP: Digital Signal Processor
- MuG: Multiple Gate

1 Introduction

1.1 CMOS technology

We live in an era where technology has a major role in our daily lives. From the phone we have in our pockets or the computer we work with to the cars we drive, the photos we take, or even the medical equipment improving our life quality (pacemakers, cochlear implants,...) The majority of these technological devices have a chip that controls their functionality. And the primary building block of those chips is the transistor.

The first transistor was invented in 1947 by Bardeen, Brattain and Shockley, conceived to replace vacuum valves. Not long after that milestone, in 1958, the first integrated circuit (IC) was developed by Jack Kilby. This chip integrated 6 transistors into the same semiconductor base, made of germanium. Parallelly, Robert Noyce developed a method to create metal interconnections for building an IC with many components, introducing the "planar process" in 1959 [3].

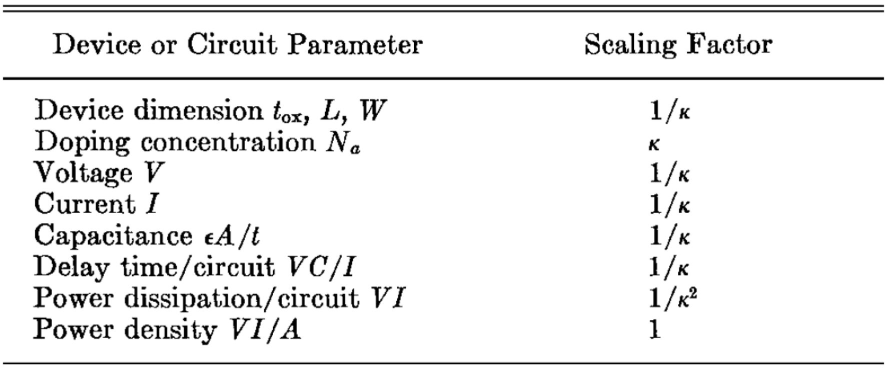
Transistors, while initially accounted in cm dimensions, rapidly scaled to mm size and then to µm, enabling smaller electronic appliances and paving the way to chip miniaturization. This ability to shrink transistors allowed an astonishing decrease in fabrication unit cost. As its size diminished, more transistors could fit in a silicon wafer, increasing the number of circuits per wafer, hence reducing significantly the cost per chip. That reduction in cost boosted the development of this technology.

The progressive shrinkage of transistors led Gordon Moore to postulate in 1965 the widely known "Moore's Law" [12]. He stated that the size of transistors would shrink every two years so that the number of transistors per wafer would increase by a factor of two. This empirical law set the roadmap for the evolution of the semiconductor industry for many decades.

The process of miniaturizing a transistor consists fundamentally on reducing the length of the gate L. Every technology node is associated to a different length value, significantly decreased with each new generation. Examples of these technology nodes are 0.18 µm, 0.13 µm, 90 nm, 65 nm... [6]

Besides gate length, some other parameters are reduced when transistors are shrunk. In 1974, Robert H. Dennard published a consistent set of scaling relationships between design parameters for the miniaturization of CMOS devices [5]. There, he stated that as the device dimensions scale by a factor of $1/\kappa$, parameters such as power supply voltage V_{dd}, current I and capacitance C must also be reduced by that factor (Table 1.1). As a consequence, the power dissipation of each circuit is reduced by $1/\kappa^2$ due to the reduced

Table 1.1: Scaling relationships for device parameters [5].



Device or Circuit Parameter	Scaling Factor
Device dimension t_{ox}, L, W	$1/\kappa$
Doping concentration N_a	κ
Voltage V	$1/\kappa$
Current I	$1/\kappa$
Capacitance $\epsilon A/t$	$1/\kappa$
Delay time/circuit VC/I	$1/\kappa$
Power dissipation/circuit VI	$1/\kappa^2$
Power density VI/A	1

voltage and current levels. Given that the area of a given device or circuit is also reduced by $1/\kappa^2$, the power density remains constant.

Thanks to that, manufacturers were able to drastically raise frequencies from one generation to the next without significantly increasing circuit power consumption, achieving notable reduction of circuit delay, as shown in Fig. 1.1. That way, integrated circuit speed increased roughly 30% with each new technology node, leading to much more involved functionality [4].

1.2 Nano-scale CMOS technology problems

As previously explained, strictly following Dennard's scaling rules kept the power density constant. This miniaturization method is known as constant field scaling, because the electric field inside the FET is unaltered by the reduction.

However, early generations of MOSFET did not follow this trend for scaling voltages. Instead, they maintained them high to obtain stronger electric fields that would increase carriers velocities, yielding higher transistor performance. High-field effects were also kept in check by the gradually descending voltage [7].

That way, the power could not be kept constant and instead increased dramatically each technological node, as it can be clearly noted from data published in the International Solid-State Circuits Conference (ISSCC) in 2001 [16] (see Fig. 1.2).

That did not pose a problem for micro scale-technologies, but as transistors started to reduce further in size, eventually reaching nano dimensions (beyond 90 nm [24]), excessive power consumption became a critical issue.

The most important effect that came along with increased power was heat dissipation [23]. As chips got denser, power consumption began to put circuits at risk of overheating [19]. An increase of the chip temperature leads to fundamental problems, due to the direct

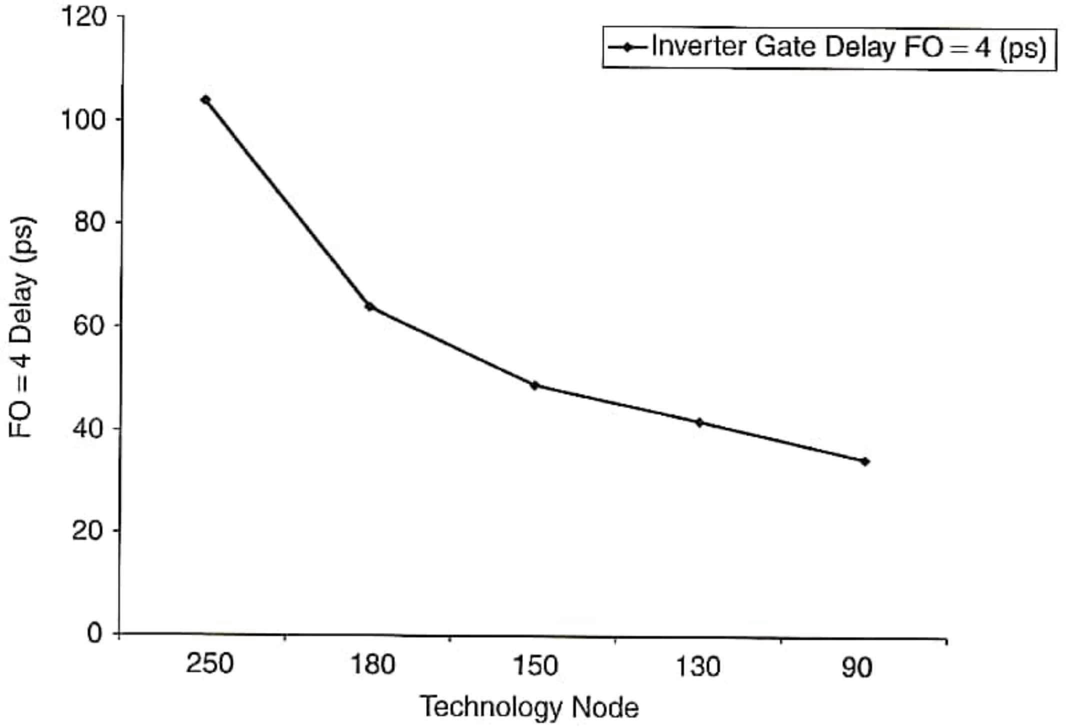


Figure 1.1: Gate delay versus technology node (in nm) [24].

influence of heat in transistor operation. High temperatures can eventually lead to chip malfunctioning, non-modeled behaviours or even irreversible physical damages.

With billions of transistors being placed on an integrated circuit, power dissipation became a fundamental problem [11]. That way, nanoelectronic technologies started to focus on how to reduce power consumption.

1.3 Reduction of power dissipation

Digital circuits are the dominant circuitry in nowadays technology. Moreover, they are often taken as the reference in metrics like delay (Fig. 1.1) or power consumption (Fig. 1.2). Therefore, the main focus of this study is concentrated on digital circuits.

Broadly speaking, power dissipation in digital circuits can be divided into two different components: dynamic power and static power [1].

Dynamic power is the power consumed while the circuit is operating in on-state. It is:

$$P_{on} = V_{dd} \, I_{on} \ . \tag{1.1}$$

Static power is the one dissipated when the device is in off-state. Its main source are

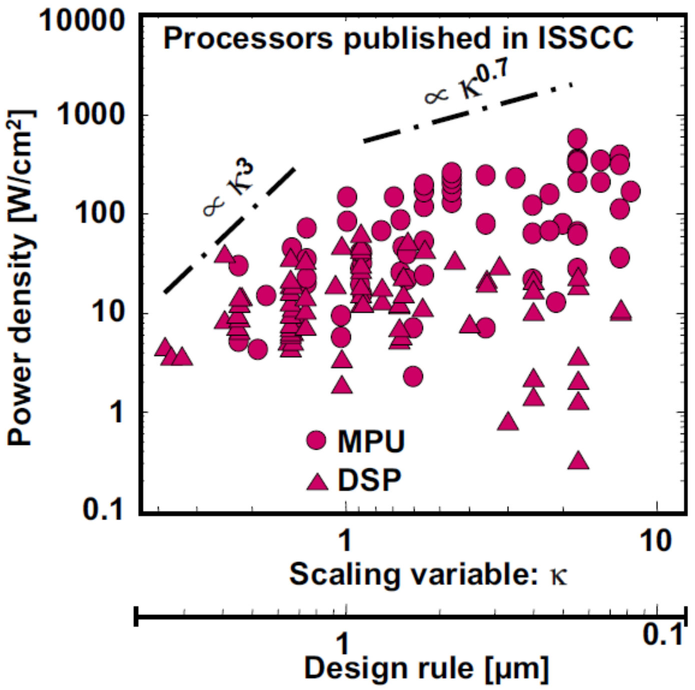


Figure 1.2: Power density increase due to scaling in processors [16].

leakage currents I_{off}:

$$P_{off} = V_{dd}\ I_{off}\ . \tag{1.2}$$

Note that transistors in digital circuits mainly operate by switching between "on" and "off", due to the discrete nature of digital signals. Consequently, the gate voltage takes the values $V_G = V_{dd}$ in on-state and $V_G = 0$ in off-state to exploit the whole voltage range.

In nano-scale traditional bulk CMOS technologies, six leakage mechanisms contribute to total static power dissipation, e.g., tunneling current through gate, reverse junction bias current. The dominant one is the subthreshold current [14] and it can be clearly noted from Fig. 1.3. Moreover, its effect is expected to increase more than the rest of leakage components with scaling [1], as it is shown in Fig. 1.4. Therefore, subthreshold current is the leakage current of greater importance when considering modern nano-CMOS devices

$$I_{off} \approx I_{DS,\ \text{subthreshold}}\ . \tag{1.3}$$

Subthreshold current is the current that flows through the channel when the device is in off-state, that is, when the gate voltage is lower than the threshold voltage $V_G < V_T$ (for an NMOS). As the gate voltage rises from 0 to V_T, charge carriers accumulate in the channel gradually until reaching a substantial amount for $V_G = V_T$, the point where the transistor "turns on" and the device is considered to enter strong inversion. For gate voltages below V_T, the charge density approaches zero rapidly. But on a logarithmic scale, for several

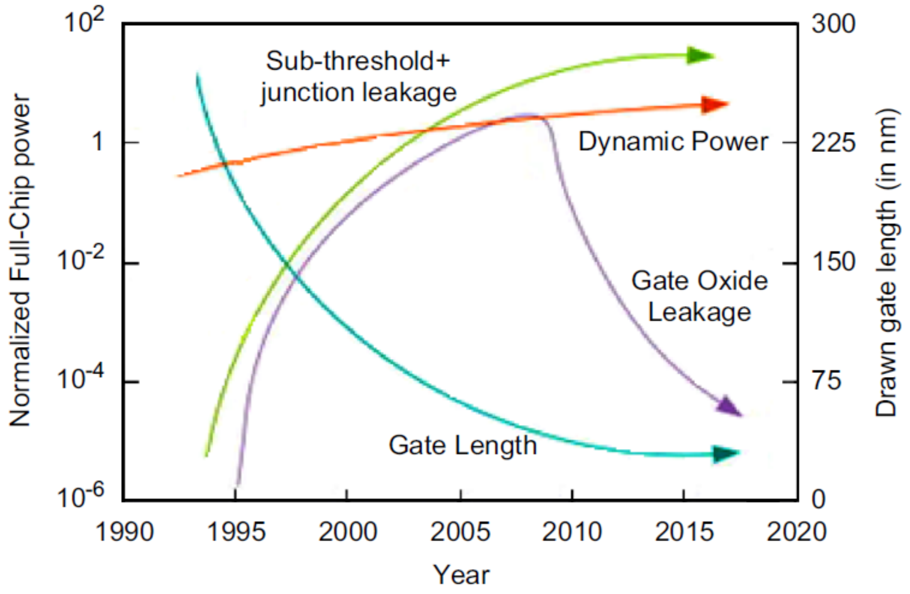


Figure 1.3: Trends of major sources of power dissipation in nano-CMOS transistor [1].

tenths of millivolts below V_T, the amount of charge in the channel is non-negligible [22]. That mobile charge density below threshold is the one that gives rise to the subthreshold current.

Subthreshold current will be studied in this work deeply, but to understand its importance in relation with power consumption, its exponential dependence with the so called overdrive factor $(V_G - V_T)$ should be stated in advance:

$$I_{DS,\ \text{subthreshold}} \sim e^{q(V_G - V_T)/mkT} . \tag{1.4}$$

Nevertheless, an expression for the subthreshold current will be properly obtained later on.

The main problem that nano CMOS technologies face is reducing power dissipation but without decreasing circuit performance. Performance depends fundamentally on the speed of transistors to switch from on- to off-state. In order not to reduce speed, I_{on} has to be kept high so that the capacitors in the circuit can be charged and discharged quickly [11]. Therefore, the on-current cannot be decreased with scaling. That only leaves one option to reduce power consumption (Eq. (1.1)): reducing the supply voltage V_{dd}.

The on-current I_{on} depends approximately linearly on the overdrive factor,

$$I_{on} \sim (V_G - V_T)^2 , \tag{1.5}$$

so if we reduce V_{dd}, which is directly linked to V_G, we have to reduce equally V_T. That way, we ensure that we keep the overdrive factor, and hence the I_{on} current, high.

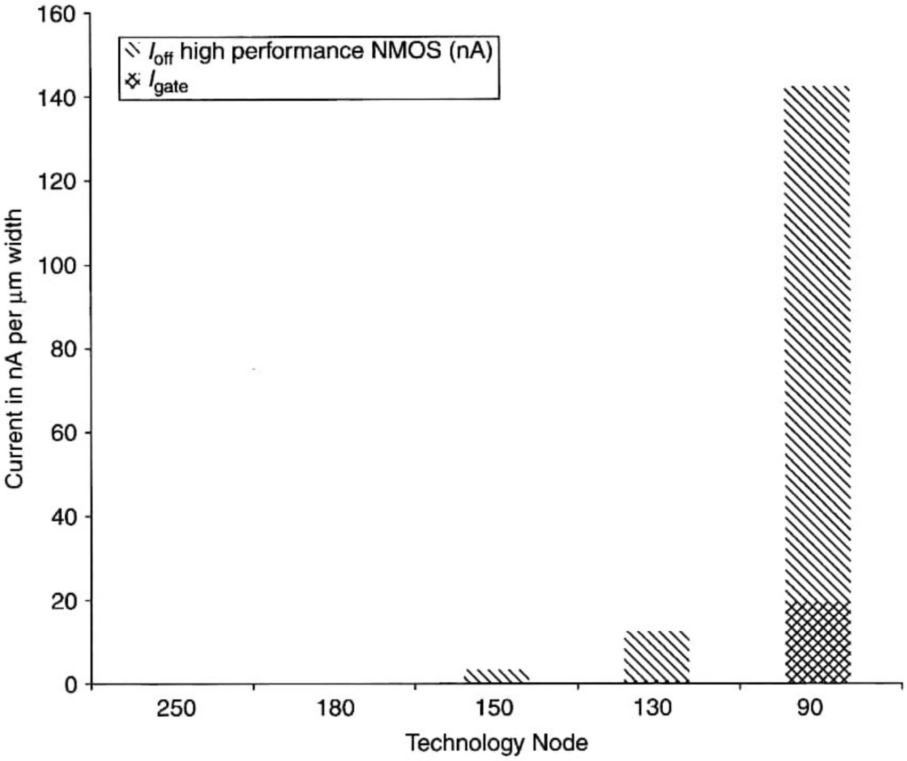


Figure 1.4: I_{gate} and subthreshold leakage versus technology node (in nm) [24].

That necessity has an enormous disadvantage emerging from the dependence of I_{off} with the overdrive factor (Eq. (1.4)). If we keep the voltage difference high, the value of I_{off} will increase exponentially, and so will the static power dissipation (Eq. (1.2)). That would lead to not diminishing the total power consumption. Therefore, reducing V_{dd} does not solve the power dissipation problem.

In Fig. 1.5, the exponential increase of I_{off} when decreasing V_{dd} can be observed. When the supply voltage is decreased along with V_T (the difference pointed by the red arrow remains constant), the NMOS current characteristic moves to the left. That way, I_{off} (the intersection with the drain current axis) increases.

If we could decrease V_{dd} and V_T without further increasing I_{off}, we could achieve a significant reduction in power dissipation without any reduction in speed. In order to succeed in that, we would need to increase the slope of the characteristic in the subthreshold region (below V_T).

The inverse of the slope S in Fig. 1.5 is a magnitude known as subthreshold slope or subthreshold swing SS [11]:

$$SS = \frac{1}{S} = \left\{ \frac{d\left[\log_{10}(I_{DS})\right]}{dV_G} \right\}^{-1} . \tag{1.6}$$

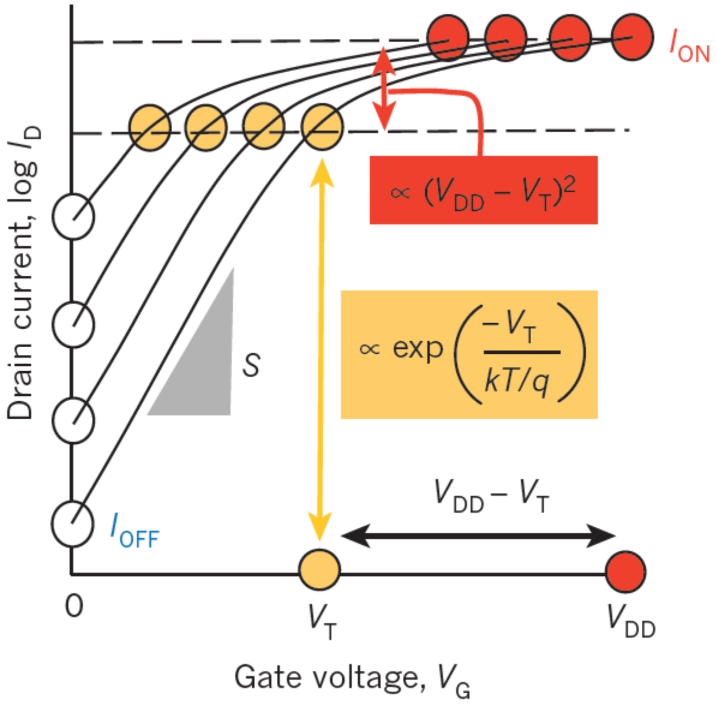


Figure 1.5: Transfer characteristics (I_{DS} vs. V_G) of a MOSFET showing an exponential increase in I_{off} because of the inherent limitation of SS [10].

This parameter tells us how much gate voltage needs to be applied in order to increase the drain subthreshold current by a factor of 10. The smaller the SS, the lower the voltage we need to increase I_{off} one order of magnitude.

The subthreshold slope does not scale with the rest of device parameters as the transistor is miniaturized. In fact it has an infrangible lower limit of 60 mV/decade [10]. Therefore, the characteristic's slope cannot be modified to reduce power dissipation without loosing performance.

But why? What is the physics behind the impossibility of scaling the subthreshold slope? In this work we aim to study the physical principles behind that lower bound in order to fully understand where that limitation stems from. With those answers, we will be able to see if there is any way to solve the power dissipation problem in CMOS technologies, and if not, to seek for new devices with different physical properties governing their functioning in the right direction, so that the problems that nowadays technologies face can be overcome.

2 Theoretical background

2.1 Basics of MOSFETs

2.1.1 Physical structure

The Metal-Oxide Semiconductor Field-Effect Transistor (MOSFET) has been the most used transistor in the field of microelectronics for many years. These transistors, also called MOS devices, are transistors with 4 terminals: gate (G), bulk (B), source (S) and drain (D).

The gate is the top conductive plate that lies on a thin dielectric layer, which is deposited on the underlying silicon substrate.

The silicon substrate, also known as bulk or body, can be p-type or n-type silicon. Depending on the type of doping, it gives rise to one of the two flavors of MOSFET: NMOS (p-type substrate) or PMOS (n-type substrate). The one depicted in Fig. 2.1 is an NMOS. It should be highlighted that typically the bulk terminal is not represented, because it is normally connected to ground.

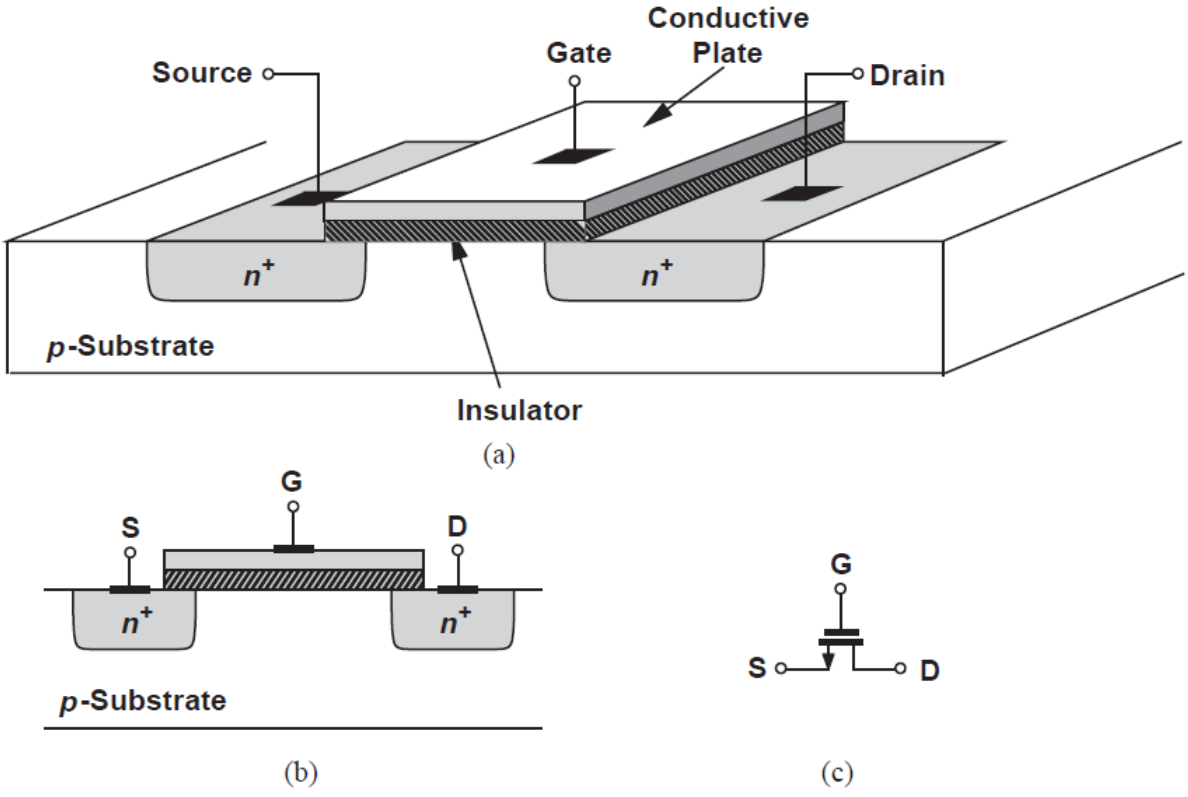


Figure 2.1: (a) Structure of a MOSFET, (b) side view, (c) circuit symbol [15].

The remaining two terminals are two heavily-doped regions in the substrate. These two highly-doped regions, n+ for NMOS or p+ for PMOS, are called source (S) and drain (D) to indicate that the former can provide charge carriers and the latter can absorb them [1].

The MOS device is symmetric with respect to S and D because the doping concentration is nominally the same for both terminals. Therefore, the identification of S and D terminals depends on a voltage convention ($V_S < V_D$) instead of on different physical or electrical properties. This symmetry differentiates the MOS transistor from others, such as the Bipolar Junction Transistor (BJT).

The materials employed to build the device are generally standardized. The gate plate must serve as a good conductor and was in fact realized by metal (aluminum) in the early generations of MOS technology. However, it was discovered that noncrystalline silicon (polysilicon), with heavy doping to achieve low resistivity, exhibits better fabrication and physical properties [15]. Therefore, it is the one used nowadays to manufacture the gate.

The dielectric layer in between the gate and the substrate plays a critical role in the performance of transistors and is created by thermally growing a thin layer of silicon dioxide SiO_2 on top of the substrate area.

2.1.2 MOS operation

The operation of the MOS transistor is governed by the different voltage conditions of the gate, source and drain terminals. For simplicity, only a NMOS transistor will be analyzed, yet the functioning of a PMOS is analogous under complementary voltage conditions.

The aim of this section is to give an overview of the MOSFET operation. To fully understand this operation, a deeper electrostatic analysis should be carried out. In this work, such analysis will be done for the particular region of interest, i.e., subthreshold region. Still, for the sake of a bigger picture, an explanation of the global functioning is provided here.

To begin, source and drain terminals will be considered to be grounded and gate voltage will be increased from 0 V. Note that between source and bulk, as well as between drain and bulk, there are two pn-junctions (see Fig. 2.1) that need to be reverse biased, in order not to have an undesired flow of charge through the device. Therefore, V_S and V_D (i.e., the source and drain voltages, respectively) must be always larger than the bulk's voltage. Consequently, in this case, the substrate will be grounded too.

As illustrated in Fig. 2.2, the gate voltage increases from 0 to a positive value, a positive

[1] Depending on the flavour of the transistor (NMOS or PMOS) this emitted and collected charge carriers are electrons and holes, respectively.

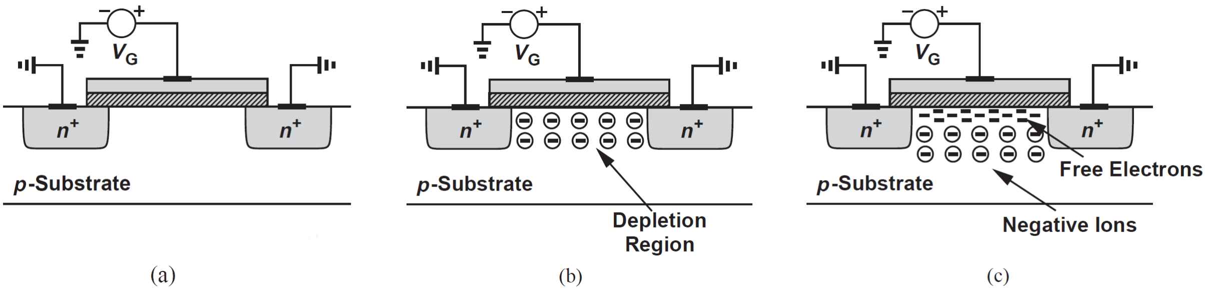


Figure 2.2: Process of channel creation as the gate voltage increases (from left to right). Modified figure from [15].

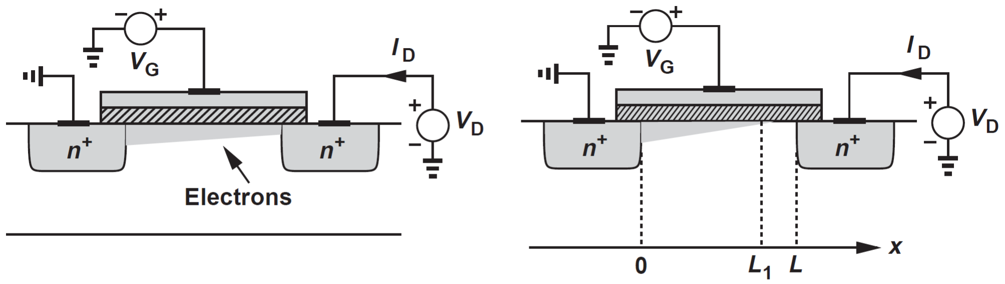
charge density is effectively accumulated in the polysilicon plate. To mirror this effect, a negative charge density will appear in the channel. Therefore, holes are repelled leaving behind ionized acceptor atoms. The region near the interface with the oxide acquires then a negative charge density. But acceptor ions are not free to move through the semiconductor and therefore they do not form an effective channel.

As V_G keeps increasing, free electrons are attracted to the interface with the oxide (Fig. 2.2 (c)). These electrons are provided by the n+ source and drain regions, and not by the substrate [15]. They are mobile charge carriers, but yet no current flows through the device, as they remain along the interface with the oxide with zero net motion. Note that as the gate terminal is separated from the channel by an insulator layer, no current will flow through the gate.

When the electron concentration becomes sufficiently large, the device is considered to be in on-state. The gate voltage at which it"turns on" is called threshold voltage V_T. Note that the point at which the device turns on is chosen by convention to be when the mobile electron concentration is the same as the hole concentration (majority carriers) at any other point (non-altered by voltage conditions) of the p-substrate. This is an arbitrary choice, and the accumulation of charge carriers in the channel is in fact a gradual process. Therefore, when the device is in off-state, there are also electrons in the channel. That non-zero electron concentration is what gives rise to the subthreshold current that was previously mentioned.

Once the channel has been created, in order to have a current flow, we need to increase the drain voltage. If the drain voltage is then higher than the source voltage, the potential at each point along the channel with respect to ground increases as we go from the source towards the drain. This effect arises from the gradual voltage drop along the channel due to the resistive behaviour associated to semiconductors. The density of electrons in the channel follows the same trend, falling to a minimum at the end of the channel, next to the drain (see Fig. 2.3).

When V_D is slightly increased, electrons start to move from source to drain creating a



$V_D < V_G - V_T$

(a)

$V_D > V_G - V_T$

(b)

Figure 2.3: Graphical representation of the channel in (a) linear, (b) saturation regions. Modified figure from [15].

current flow in the opposite direction. At that state, the device operates as a voltage controlled resistor as a result of the gate voltage controlling the amount of charge carriers available to conduct current. This operation region is known as linear or ohmic region.

If V_D keeps increasing, it will eventually surpass the value of $V_G - V_T$. If that occurs, the end of the channel $x = L$ will not have enough potential difference to attract electrons, and the channel will cease to exist at that point. This effect is known as channel pinch off. If V_D further increases, this pinch off point will be displaced towards the source, reducing the effective channel length to $L_1 < L$, as depicted in Fig. 2.3 (b).

As a consequence, the device has no channel in between L_1 and L. Despite that, it still conducts current because when the electrons reach the end of the channel, they are strongly attracted by the electric field of the bulk-drain depletion region and they are drifted to the drain terminal. Nonetheless, at this state the drain voltage no longer affects the current significantly, and the MOSFET behaves as a current source. This region of operation is known as saturation region, because the current is saturated as it does not further increase with the drain voltage (neglecting second-order effects, e.g., channel length modulation).

Summarizing, when the device is in on-state ($V_G > V_T$), as V_D increases, three regions of operation are distinguished: cut-off ($V_D = 0$), linear region ($V_D < V_G - V_T$) and saturation ($V_D > V_G - V_T$). When the gate voltage does not exceed the threshold value ($V_G > V_T$), the device is in subthreshold region.

All in all, the MOS transistor is a device that permits its current to be easily controlled and modified by other device parameters, and its dependence upon them changes under varying biasing conditions.

2.2 Gate voltage and surface potential

When a voltage is applied to the gate electrode, not all of it reaches the semiconductor surface. The presence of the insulating layer of silicon oxide separating the gate from the channel results in a fraction of the gate voltage dropping across the SiO_2. That way, the voltage is distributed across the whole MOS structure. As a consequence, the effective voltage across the channel is the gate voltage minus the voltage drop across the oxide. Mathematically, that is:

$$V_G = V_{ox} + V_{SC} \ . \tag{2.1}$$

Assuming that the voltage drop at the ohmic contact of the substrate is negligible, and that the bulk is grounded, all the potential that reaches the semiconductor surface will drop across its body. Therefore, the total potential drop across the semiconductor is exactly the value of the potential at the surface ψ_s:

$$V_{SC} = \psi_s \ . \tag{2.2}$$

2.3 Capacitive MOS structure

The Metal-Oxide-Semiconductor can be seen as a capacitive structure analogous to a parallel-plate capacitor, a passive component formed by two metal plates separated by an insulator layer. It is well known that its capacitance per unit area is given by:

$$\frac{C}{A} = \frac{\epsilon_{ins}}{t_{ins}} \quad [\mathrm{F/m^2}] \ , \tag{2.3}$$

where ϵ_{ins} is the permittivity and t_{ins} the thickness, both of the insulating layer.

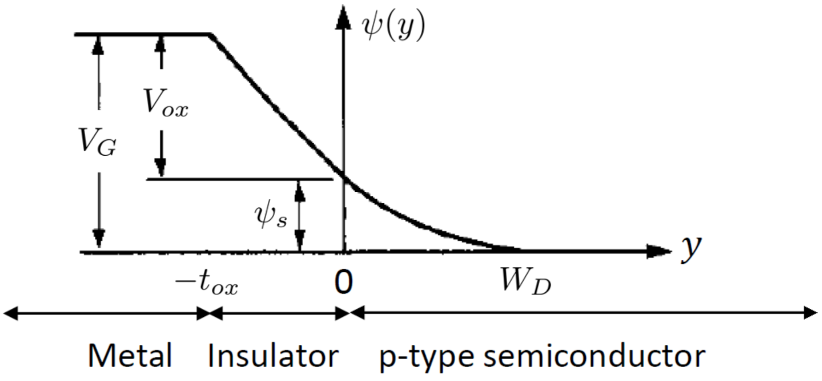


Figure 2.4: Voltage across the MOS structure. Modified figure from [20].

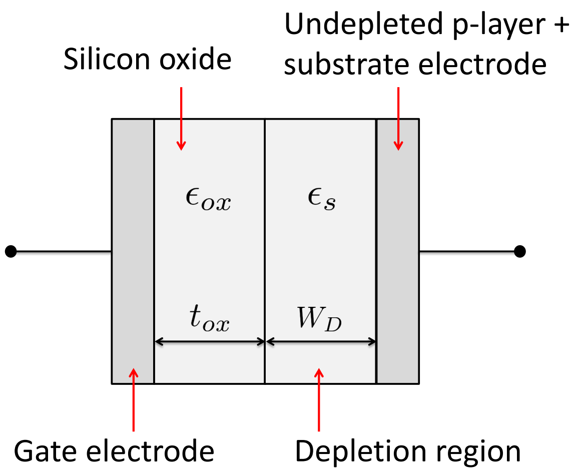


Figure 2.5: Capacitive equivalent of the MOS structure.

The MOS transistor, instead of two metal plates, has the polysilicon gate and the semiconductor substrate playing the role of charge accumulators, separated by the insulating SiO_2 layer.

As we previously outlined, when a positive voltage is applied to the gate electrode, a positive charge density is effectively deposited on it. In response, an equal net negative charge is accumulated in the semiconductor. As V_G increases from zero to a positive value, the acceptor atoms near the surface of the semiconductor start to acquire electrons to compensate the positive charge in the gate, thereby becoming negative ions. This acceptors are not distributed on the surface of the semiconductor, like electrons would be in a metal plate. Instead, they are spread out within a certain volume adjacent to the interface with the oxide. That volume is, in fact, a depletion region of width W_D. As a result, the surface potential ψ_s decreases along the distance W_D in the semiconductor material, becoming zero outside this region, as shown in Fig. 2.4. Note that the resistive behaviour of the undepleted part of the p-type semiconductor is neglected, therefore no voltage drops between the end of the depletion region and the bulk electrode.

At the same time that the depletion region arises, free electrons are attracted to the semiconductor surface initially creating a reduced density of free charges. When the surface potential is not high, as in subthreshold region, the amount of free electrons is very small compared to the depletion ion charges.

That way, in subthreshold region, the MOS structure can be modelled by a capacitive structure with two different dielectrics in series [11]. The first capacitor's plate would correspond to the gate electrode, as depicted in Fig. 2.5. The two dielectric materials would be, first the silicon oxide, and second the depleted region of the semiconductor. As we previously stated, the amount of free electrons at the surface in the subthreshold region is negligible compared with the depletion charges. Therefore, the depletion region

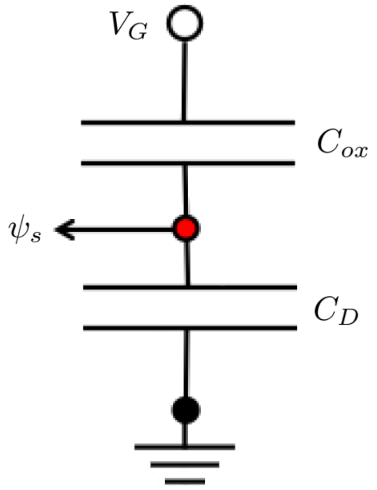


Figure 2.6: Equivalent circuit of the MOS structure. Modified figure from [11].

is effectively an insulator, due to the lack of mobile charges. Consequently, the other plate would correspond to the undepleted p-layer and the substrate contact.

The capacitance of a capacitor made up of two different dielectric materials in series, with different dielectric constants and thicknesses, is given by the sum of the inverse of the capacitances. Therefore, the total gate capacitance C_G of the MOS structure would be:

$$\frac{1}{C_G} = \frac{1}{C_{ox}} + \frac{1}{C_D} \ . \tag{2.4}$$

The oxide and depletion capacitances C_{ox} and C_D are also expressed in terms of a parallel-plate capacitor (Eq. (2.3)). Their values are, respectively,

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} \ , \tag{2.5}$$

$$C_D = \frac{\epsilon_s}{W_D} \ . \tag{2.6}$$

The equivalent circuit of the MOS structure in subthreshold region is depicted in Fig. 2.6. Applying charge conservation to this circuit yields:

$$\psi_s = V_G \left(\frac{C_{ox}}{C_{ox} + C_D}\right) = \frac{V_G}{m} \ , \tag{2.7}$$

where m is known as the body effect coefficient in depletion

$$m = 1 + \frac{C_D}{C_{ox}} \ . \tag{2.8}$$

The body effect coefficient m tells us the fraction of the applied gate voltage that drops across the semiconductor.

3 Physics of the subthreshold region

After presenting the importance of subthreshold current as one of the main sources of leakage and hence power dissipation, and outlining the most important aspects of the MOSFET operation, in this section the physics of subthreshold region is going to be studied in depth. The main objective is to obtain an analytical expression for both the subthreshold current and the subthreshold slope so that we are able to gain insight into the different physical principles governing them.

Our analysis is going to be focused on a NMOS transistor, yet for a PMOS an analogous approach can be followed.

To obtain an expression for the drain to source subthreshold current, first the transport process of carriers needs to be studied. That will be done for different biasing conditions, starting from no voltage applied to any of the terminals, following with an increase of the gate potential, and finishing by incorporating a positive voltage in the drain region.

3.1 No applied voltage ($V_G = 0,\ V_S = V_D = 0$)

In the transistor three zones of different doping are found: the source, the channel, and the drain. As previously stated, the source and drain regions are heavily doped with donor atoms (n-type). Conversely, the channel is made up of p-type silicon. As illustrated in Fig. 3.1, in a uniformly doped bulk semiconductor, the energy bands are independent of the position. For n-type materials, the Fermi level lies near the conduction band E_c, whereas for a p-type it lies closer to the valence band E_v.

When the three zones are brought together, as the Fermi Level must be constant for the joint material, the energy bands have to bend. To line up the Fermi levels in the three regions, source and drain energy bands must drop or, equivalently, the channel must rise in energy, until E_F is constant. Physically, the alignment of the Fermi levels occurs because electrons flow from source and drain regions to the channel, which sets up a charge imbalance and produces an electrostatic potential difference between the adjoined regions. As in a p-n junction, this process originates a depletion region in both interfaces. The junctions between source and channel and between channel and drain are assumed to be step junctions in this analysis.

The electrostatic potential arisen by the charges displacement is known as the built-in potential V_{bi}. It determines the height of the energy barrier originated along the channel

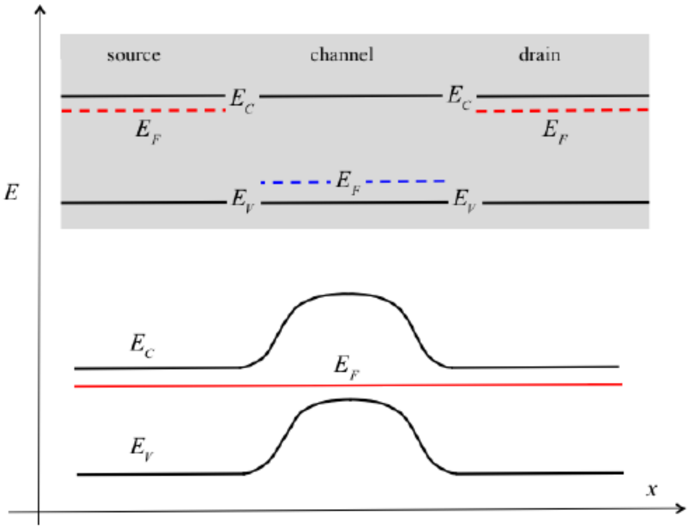


Figure 3.1: Diagram of the equilibrium energy bands. Top: separately source, channel, and drain regions. Bottom: The three regions after metallurgical junction with $V_S = V_D = V_G = 0$ [11].

from the bands bending. Its expression, known from basic semiconductor theory, is [20]:

$$V_{bi} = \frac{kT}{q} \ln \frac{N_d N_a}{n_i^2} \ , \tag{3.1}$$

where k is Boltzmann's constant, T the absolute temperature, q the electron charge, N_d and N_a the donor and acceptor concentrations for the n- and p-regions, respectively, and n_i the carrier density of the intrinsic semiconductor.

To evaluate carrier transport through the three regions, one can focus on the study of one of the two p-n junctions, and as the device is symmetric, the behaviour can be extrapolated to the other junction. Having said that, we will first analyze the source-channel junction. Furthermore, as our study concentrates on the NMOS, the channel is going to be created by electrons, and so the subthreshold current. Therefore, this study will address the transport of this type of carrier, yet an analogue approach can be followed for holes in a PMOS.

In an n-type region, electrons constitute majority carriers. Assuming total ionization of donors, a heavily doped material and a negligible concentration of acceptors, at room temperature the electron concentration at equilibrium n_{0n} can be approximated as:

$$n_{0n} \simeq N_d \ . \tag{3.2}$$

For the p-region, electrons correspond to minority carriers. Assuming total ionization of acceptors, heavy doping and a negligible amount of donors, at room temperature and

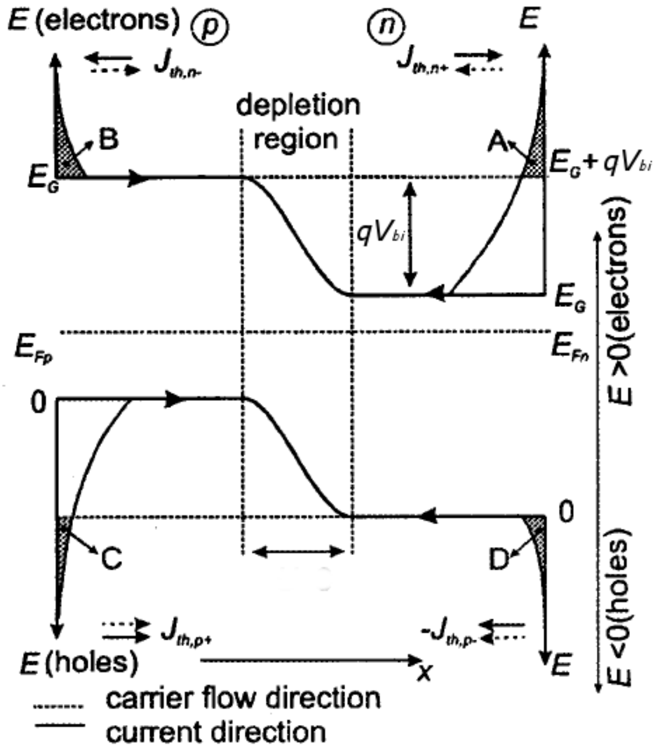


Figure 3.2: Energy band diagram of a p-n homojunction in equilibrium [2].

under equilibrium conditions the electron concentration of the p-type material n_{0p} is

$$n_{0p} \simeq \frac{n_i^2}{N_a} \ . \tag{3.3}$$

As depicted in Fig. 3.2, there is a fraction of electrons in the n-side conduction band located at energy levels above the barrier, that is, with energies $E \geq E_G + qV_{bi}$. This fraction of electrons can be obtained from the Fermi-Dirac distribution function and the density of allowed states in the conduction band, thereby presenting a direct dependence on temperature. Carriers with energies larger than the height of the barrier can be emitted and then injected across the depletion region [2]. This thermally induced flow of charge carriers over a potential-energy barrier is known as thermionic emission.

The thermionic current density flowing from n- to p-side is denoted by $J_{th,n+}$ in Fig. 3.2 because the motion of carriers follows negative x direction and, as they are electrons, the current goes in the opposite direction. Similarly, the current density flowing from p- to n-side is denoted by $J_{th,n-}$.

In equilibrium (i.e., with no external voltage applied) the fraction of electrons able to surpass the barrier is the same as the electron concentration in the p-region. In terms of energy states, that means that electrons fill the conduction bands up to a level that has the same energy in both p- and n-regions.

Therefore, n-side electrons do not have enough energy to reach the empty p-side levels. The energy levels that they could occupy are already filled by the p-side minority electrons. Same happens for electrons in the p-side that could flow to the n-side.

Consequently, the only way for electrons to change sides is through an interchange. For example, an electron from the n-side is thermionically emitted to the p-region, leaving an empty energy level behind. Then, an electron from the p-side flows to the n-side to occupy this energy level, leaving in turn a gap in the p-side conduction band. Hence, the n-side electron fills it up.

Through this type of exchange the net flow of charge carriers is zero because:

$$J_{th,n+} = J_{th,n-} \ . \tag{3.4}$$

Thus, the net current density in equilibrium is zero too:

$$J_{th} = J_{th,n+} - J_{th,n-} = 0 \ . \tag{3.5}$$

3.2 Positive gate voltage ($V_G > 0,\ V_S = V_D = 0$)

Now biasing conditions are changed by increasing the gate voltage from 0 to a positive value. The gate voltage V_G, and thereby the surface potential ψ_s, affects the energy bands, and its effect is going to be analyzed separately in the x- and y-directions.

3.2.1 Analysis along y-direction

When a positive voltage is applied to the gate, and hence to the semiconductor, a downward bending in the energy bands occurs. We expect such a tilt since the applied electric field carries a potential energy that contributes to the total energy of the electrons in the material, therefore modifying the energy band diagram. The electric field causes a gradient in E_i and similarly in E_v and E_c, that can be expressed as [20]:

$$\xi(y) = -\frac{d\psi(y)}{dy} = -\frac{d}{dy}\left[\frac{E_i}{(-q)}\right] = \frac{1}{q}\frac{dE_i}{dy} \ . \tag{3.6}$$

As previously explained, the value of the potential at the surface decreases as one deepens in the p-material along y-direction, making the voltage across the semiconductor a function of the depth $\psi(y)$. Likewise, the band bending decreases progressively in the same direction along with the potential, as illustrated in Fig. 3.3.

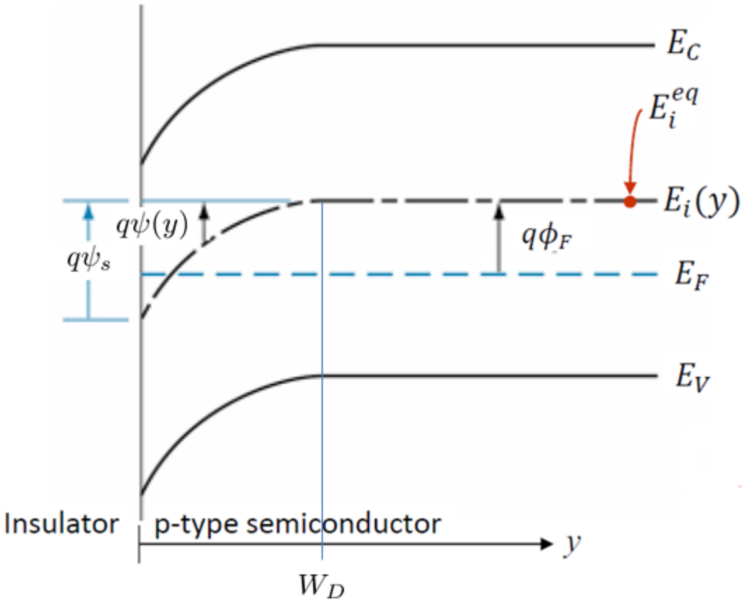


Figure 3.3: Energy band diagram of the channel in the y-direction under a positive gate voltage. Modified figure from [20].

The magnitude $q\phi_F$ in Fig. 3.3 measures the position of the Fermi level below the intrinsic level in equilibrium E_i^{eq} and indicates how strongly doped p-type the semiconductor is [20],

$$q\phi_F = E_i^{eq} - E_F \ . \tag{3.7}$$

In the region of the p-type semiconductor affected by the electric field, the potential $\psi(y)$ can be expressed as:

$$q\psi(y) = q\phi_F - [E_i(y) - E_F] \ , \tag{3.8}$$

where $E_i(y)$ is the bent intrinsic energy band as a function of the depth y.

From basic semiconductor theory, we know that the equilibrium electron concentration for both p- and n-type semiconductors is [20]

$$n_0 = n_i \ e^{-(E_F - E_i^{eq})/kT} = n_i \ e^{-q\phi_F/kT} \ , \tag{3.9}$$

and this expression can be written in terms of $q\phi_F$ using the previous definition of that magnitude (Eq. (3.7)).

The electron concentration at any point y can be easily related, as done in [20], to Eq. (3.9). The resulting expression can be analogously presented in terms of ϕ_F and $\psi(y)$ using Eq. (3.8), and further simplified with Eq.(3.9) yielding

$$n(y) = n_i \ e^{-[E_F - E_i(y)]/kT} = n_i \ e^{-q[\phi_F - \psi(y)]/kT} = n_0 e^{q\psi(y)/kT} \ . \tag{3.10}$$

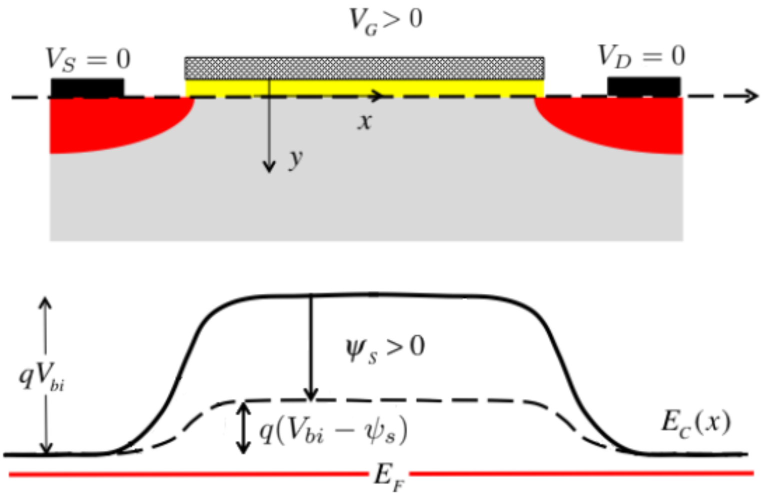


Figure 3.4: Channel barrier with a positive gate voltage applied [11].

Eq. (3.10) is the expression of the amount of electrons that the semiconductor can potentially hold due to the band bending emerging from the applied voltage. In the study of the capacitive MOS structure, neglecting drain and source terminals, those new allowed states are filled with thermally generated electrons (minority carriers) from the p-type material. Nevertheless, when the drain and source terminals are considered, those extra electrons are majority carriers from the n-regions with enough thermal energy to surpass the barrier by a thermionic emission process and flow to the p-type channel.

3.2.2 Analysis along x-direction

Focusing on the x-direction, the application of a positive gate voltage causes a uniform lowering of the conduction band along the whole p-type channel. The amount of energy the conduction band is lowered corresponds to the value of the potential $\psi(y)$ at the surface, that is, at $y = 0$, as we can see in Figs. 3.3 and 3.4.

In the junction between the source and the channel, thermionic transport is also affected by the voltage. The amount of electrons in the n-type region with enough thermal energy to surpass the barrier is increased since its height has diminished. As the p-side conduction band has been lowered due to the gate voltage, now there are electrons in the n-side with enough energy to reach empty energy levels in the p-side conduction band. Therefore, electrons can be injected over the barrier and they increment the concentration of electrons in the p-region, near the border of the depletion region. The excess electrons $n_{p,\text{edge}}$ that arrive from the n-region have +x directed velocities. Thus, the electron current in the -x direction is determined by the concentration at the depletion region edge, and can be written as [2]:

$$J_{th,n} = q\ v_{th}\ n_{p,\text{edge}}\ . \tag{3.11}$$

Due to the increased concentration at the edge, the transport of excess electrons through the quasi-neutral p-region is characterized as a diffusion process described by a diffusion equation. Therefore, the net current across the depletion region must be equal to the diffusion current [2]:

$$J_{th,n+} - J_{th,n-} = J_{diff,n} = qD_n \frac{dn(x)}{dx}\ , \tag{3.12}$$

where D_n is the electron diffusion coefficient (cm^2/s).

This diffusion current comes ultimately from the fact that the concentration at the edge of the depletion region in the p-side is different from the electron concentration further in the channel.

Same process occurs in the other junction between the channel and the drain, so that when there is a positive gate voltage applied, an excess of electrons gets to the end of the channel adjacent to the drain.

But now when both junctions are considered, contributions from source and drain must be added. As both terminals are identical in terms of doping, their energy bands will also be indistinguishable. That means that the barrier has the exact same height in both junctions, and therefore when it is lowered, the same excess electron concentration will be accumulated at the beginning (next to the source) and at the end (next to the drain) of the channel. This implies that no diffusion current can be generated from the gradient in electron concentration along the channel because there is no difference in concentration. In case the channel is long enough so that the concentration in the middle is noticeably different from that at the edges and therefore a diffusion current could be originated, the net diffusion current will be zero because the contribution from both sides will be exactly the same but with opposite directions.

3.3 Positive drain and gate voltages ($V_G > 0,\ V_S = 0,\ V_D > 0$)

For a net diffusion current to appear, drain and source voltages need to come into play. Lets assume that our source is grounded and that a positive voltage is applied to the drain. The mathematical procedure described next can be extrapolated to the case where the source has also a positive voltage, but in that case the voltage difference between source and drain $V_{DS} = V_D - V_S$ must be considered, instead of only the drain voltage V_D.

As it is shown in Fig. 3.5, when a positive drain voltage is applied the electrostatic

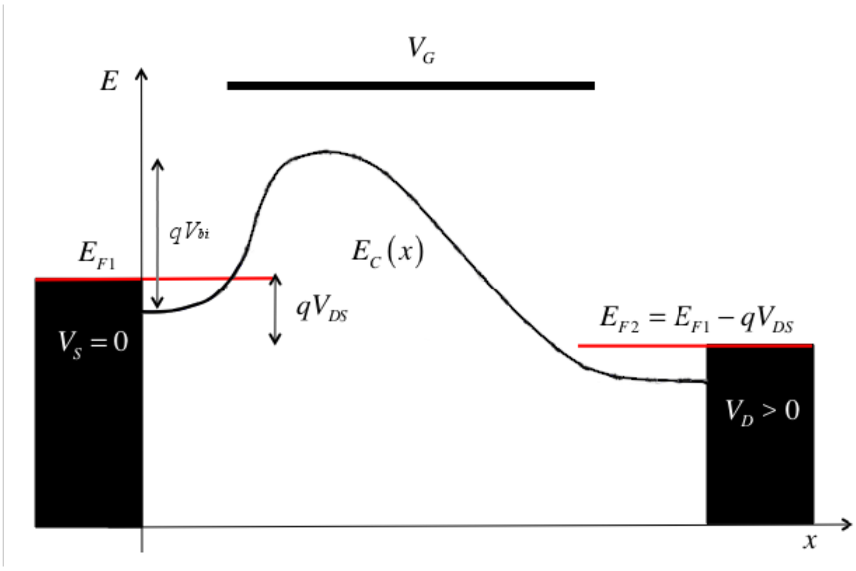


Figure 3.5: Energy band diagram in x-direction for a positive gate voltage and drain voltage. Modified figure from [11].

potential in the drain is increased and that causes a lowering of the conduction band together with the Fermi level in the n-region. The result of that lowering is an increase in the height of the potential barrier between drain and channel, which results in fewer electrons in the n-side with enough thermal energy to surpass the barrier to the channel.

Applying a drain voltage therefore causes an unbalance between the excess electron concentration at the beginning and the end of the channel. This gradient of concentration comes along with a diffusion current, that is intrinsically related to the thermionic current of electrons flowing over the potential barriers, as it was shown in Eq. (3.12). That expression can be modified to include source and drain contributions:

$$J_{Sth,n-} - J_{Sth,n+} = J_{Sdiff,n} \ , \tag{3.13}$$

$$J_{Dth,n+} - J_{Dth,n-} = J_{Ddiff,n} \ , \tag{3.14}$$

leading to:

$$J_{Sdiff,n} - J_{Ddiff,n} = J_{diff,n} = qD_n \frac{dn(x)}{dx} \ . \tag{3.15}$$

Fig. 3.6 depicts a graphical representation of all the current contributions to give a clearer picture of the process.

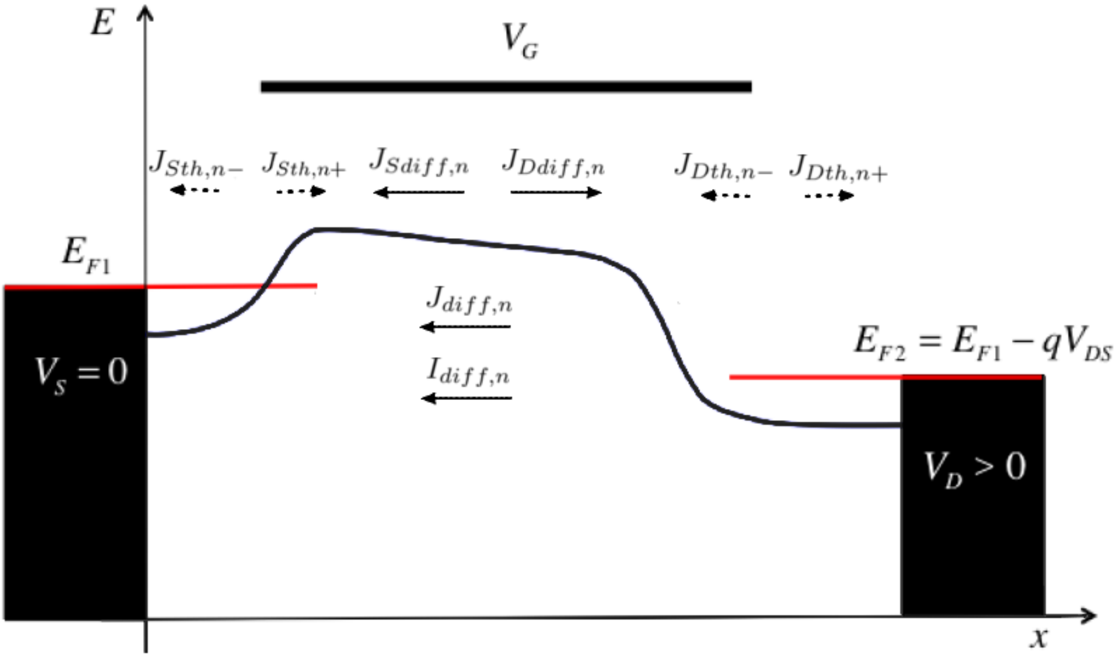


Figure 3.6: Different current density contributions. Modified figure from [11].

3.4 Drain-source current

To obtain an expression for the drain source current I_{DS}, the current density from Eq. (3.15) has to be multiplied by the area A crossed by the flow of carriers:

$$I_{diff,n} = qAD_n \frac{dn(x)}{dx} \ . \tag{3.16}$$

In the operation region of interest, subthreshold region, conduction is dominated by diffusion current [21] [22]. Thus, we can state that the drain source current is equal to that diffusion current:

$$I_{DS} = qAD_n \frac{dn(x)}{dx} \ . \tag{3.17}$$

As $J_{diff,n}$ is defined positive in the negative x-direction of Fig. 3.6, the concentration gradient has to be considered in that direction too, so:

$$\frac{dn(x)}{dx} = \frac{n(0) - n(L)}{L} \ . \tag{3.18}$$

The area A in our case is given by the width of the device W and the effective channel

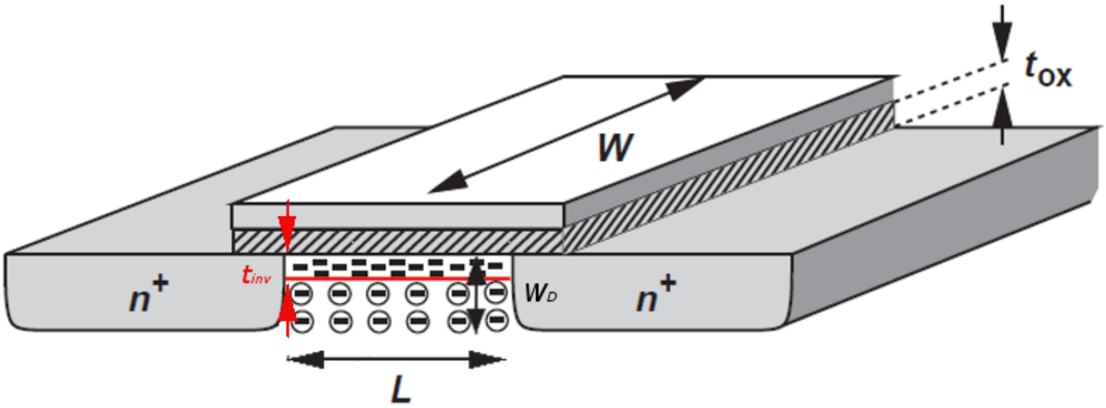


Figure 3.7: Illustration of physical dimensions of the NMOS. Modified figure from [15].

thickness t_{inv} normal to the semiconductor-insulator interface [21], as shown in Fig. 3.7,

$$A = W\ t_{inv}\ . \tag{3.19}$$

Using Einstein's relation ($D_n = \mu_n \frac{kT}{q}$) and Eqs. (3.18) and (3.19), the drain current from Eq. (3.17) can be expressed as:

$$I_{DS} = q\ \mu_n \frac{kT}{q}\ W\ t_{inv}\ \frac{n(0) - n(L)}{L}\ . \tag{3.20}$$

Because of the exponential dependence of electron density on the potential $\psi(y)$ given by Eq. (3.10), the electron layer thickness t_{inv} can be assumed to be the distance in which the surface potential decreases by kT/q [2]. Therefore, the effective channel thickness is [21]

$$t_{inv} = \frac{kT}{q\ \xi_{\text{ave}}}\ , \tag{3.21}$$

where ξ_{ave} is the average electric field in the electron layer. Given that the electron layer is thin compared to the depletion layer, the average electric field can be approximated by the value of the electric field on the surface of the semiconductor [11]:

$$\xi_{\text{ave}} \approx \xi_s\ . \tag{3.22}$$

We obtain an expression for the electric field in the y-direction by solving Poisson's equation

[2] This assumption is done by the author because at that distance, the concentration of electrons has decreased roughly a 63% with respect to the one at the surface.

with the boundary condition that the electric field has to be cancelled out outside the depletion region [11], so:

$$\xi(y) = \frac{qN_a}{\epsilon_s}(W_D - y) \ , \tag{3.23}$$

where ϵ_s is the permittivity of the semiconductor and W_D is the thickness of the depletion region.

To obtain the electric field at the surface of the semiconductor ξ_s we only need to evaluate $\xi(y)$ at $y = 0$

$$\xi(y = 0) = \xi_s = \frac{qN_a}{\epsilon_s} W_D \ . \tag{3.24}$$

Then, substituting the value of t_{inv} (Eq. (3.21)) into Eq. (3.20) all together with the value of the electric field (Eq. (3.24)), we get:

$$I_{DS} = q\ \mu_n \left(\frac{kT}{q}\right) \frac{W}{L} \frac{kT}{q\ \xi_{\text{ave}}}\ [n(0) - n(L)] \implies$$

$$I_{DS} = q\ \mu_n \left(\frac{kT}{q}\right)^2 \frac{W}{L} \frac{\epsilon_s}{qN_aW_D}\ [n(0) - n(L)] \ . \tag{3.25}$$

Besides, an expression for the electron concentration at the beginning ($x = 0$) and at the end ($x = L$) of the channel can be inferred from Eq. (3.10). As the channel thickness is quite small compared to the depletion region width, the free electron concentration can be assumed to be located along the channel at $y = 0$. Noting that the surface potential is different at the source and at the drain edges due to the existence of a positive drain potential, Eq. (3.10) yields

$$n(0) = n_{0p} e^{q\psi(x=0,y=0)/kT} = \frac{n_i^2}{N_a} e^{q\psi_s/kT} \ , \tag{3.26}$$

$$n(L) = n_{0p} e^{q\psi(x=L,y=0)/kT} = \frac{n_i^2}{N_a} e^{q(\psi_s - V_D)/kT} \ , \tag{3.27}$$

where the electron concentration in equilibrium n_{op} corresponds to the electron concentration in a p-material under no applied voltage, and is given by Eq. (3.3).

Substituting eqs. (3.26) and (3.27) into the expression of the current (Eq. (3.25)) yields

$$I_{DS} = q\ \mu_n \left(\frac{kT}{q}\right)^2 \frac{W}{L}\ \frac{\epsilon_s}{qN_aW_D}\ \frac{n_i^2}{N_a}(e^{q\psi_s/kT} - e^{q(\psi_s - V_D)/kT}) \ , \tag{3.28}$$

and taking out $e^{q\psi_s/kT}$ as a common factor, we get

$$I_{DS} = q\ \mu_n \left(\frac{kT}{q}\right)^2 \frac{W}{L}\ \frac{\epsilon_s}{qN_aW_D}\ \frac{n_i^2}{N_a}\ e^{q\psi_s/kT}(1 - e^{-qV_D/kT}) \ . \tag{3.29}$$

By definition, in strong inversion the concentration of electrons in the channel has to be the same as the concentration of holes in any other part of the p-type semiconductor, that is, it has to be equal to the concentration of acceptors

$$n(y=0) = N_a \; . \tag{3.30}$$

The value of the surface potential at which the strong inversion state is reached is

$$\psi_s = 2\phi_F = 2\,\frac{kT}{q}\ln\frac{N_a}{n_i} \; . \tag{3.31}$$

Therefore, using Eq. (3.10) and these strong inversion relations, we can express the quotient n_i^2/N_a as

$$n(y=0) = N_a = \frac{n_i^2}{N_a}\; e^{q2\phi_F/kT} \quad\Longrightarrow\quad \frac{n_i^2}{N_a} = N_a\; e^{-q2\phi_F/kT} \; . \tag{3.32}$$

Using Eq. (3.32) in the expression of the drain current (Eq. (3.29)) yields

$$I_{DS} = q\;\mu_n\left(\frac{kT}{q}\right)^2\frac{W}{L}\;\frac{\epsilon_s}{qN_aW_D}\;N_a\;e^{-q2\phi_F/kT}\;e^{q\psi_s/kT}(1-e^{-qV_D/kT}) \; , \tag{3.33}$$

and grouping the exponentials we get

$$I_{DS} = q\;\mu_n\left(\frac{kT}{q}\right)^2\frac{W}{L}\;\frac{\epsilon_s}{qN_aW_D}\;N_a\;e^{q(\psi_s-2\phi_F)/kT}\;(1-e^{-qV_D/kT}) \; , \tag{3.34}$$

so that the term qN_a can be simplified leaving the current as

$$I_{DS} = \mu_n\left(\frac{kT}{q}\right)^2\frac{W}{L}\;\frac{\epsilon_s}{W_D}\;e^{q(\psi_s-2\phi_F)/kT}\;(1-e^{-qV_D/kT}) \; . \tag{3.35}$$

From the capacitive analysis of the gate, we know that the quotient ϵ_s/W_D corresponds to the previously defined depletion capacitance (Eq. (2.6)). Introducing that into Eq. (3.35) yields

$$I_{DS} = \mu_n\left(\frac{kT}{q}\right)^2\frac{W}{L}\;C_D\;e^{q(\psi_s-2\phi_F)/kT}\;(1-e^{-qV_D/kT}) \; . \tag{3.36}$$

Then, C_D can be expressed, with the use of Eq. (2.8), as a function of the body effect coefficient m and the oxide capacitance C_{ox}, leading to

$$I_{DS} = \mu_n\left(\frac{kT}{q}\right)^2\frac{W}{L}\;C_{ox}(m-1)\;e^{q(\psi_s-2\phi_F)/kT}\;(1-e^{-qV_D/kT}) \; . \tag{3.37}$$

As the surface potential value is not directly controlled, it is more convenient to express Eq. (3.37) in terms of the gate voltage, using Eq. (2.7). Also, the factor $2\phi_F$ corresponds, as previously stated, to the value of the surface potential ψ_s needed to enter strong inversion. The threshold voltage is defined as the gate voltage needed so that a significant amount of free charge begins to build up in the channel and the transistor can be considered to be in strong inversion region. Therefore, with Eq. (2.7), the term $2\phi_F$ can also be expressed as

$$2\phi_F = V_T/m \; . \tag{3.38}$$

Considering those two relations, the exponential term in Eq. (3.37) turns into

$$I_{DS} = \mu_n \left(\frac{kT}{q}\right)^2 \frac{W}{L} \, C_{ox}(m-1) \; e^{q(V_G - V_T)/mkT} \; (1 - e^{-qV_D/kT}) \; . \tag{3.39}$$

All in all, we have obtained from a diffusion current equation an expression for the drain-source current in subthreshold region step by step, procedure that has not been reported in the literature up to now. This analysis proves that the drain-source current in subthreshold region is, in fact, a diffusion current that depends exponentially on the gate voltage. The amount of current through the channel depends directly on the fraction of electrons that are able to surpass the potential barrier, showing clearly the thermionic nature of the subthreshold current.

3.5 Subthreshold slope

Our final goal is to obtain from the subthreshold current in Eq. (3.39) an analytical expression for the previously introduced subthreshold swing SS. This parameter, also known as subthreshold slope, was defined as the change in gate voltage needed to increase the current by one order of magnitude, as well as as the inverse of the transfer characteristic slope in subthreshold region. Mathematically:

$$SS = \left\{ \frac{d\left[\log_{10}(I_{DS})\right]}{dV_G} \right\}^{-1} \; . \tag{3.40}$$

To obtain an analytical expression for SS, we take logarithm to the equation of the drain current (Eq. (3.39)) and we derive it with respect to the gate voltage V_G:

$$\frac{d\left[\log_{10}(I_{DS})\right]}{dV_G} = \frac{\mu_n \left(\frac{kT}{q}\right)^2 \frac{W}{L} \, C_{ox}(m-1) \; (1 - e^{-qV_D/kT}) \; \frac{q}{mkT} \; e^{q(V_G - V_T)/mkT}}{\mu_n \left(\frac{kT}{q}\right)^2 \frac{W}{L} \, C_{ox}(m-1) \; (1 - e^{-qV_D/kT}) \; e^{q(V_G - V_T)/mkT}} \frac{1}{\ln(10)} \; . \tag{3.41}$$

Simplifying equal terms in the numerator and denominator, we get

$$\frac{d\left[\log_{10}(I_{DS})\right]}{dV_G} = \frac{q}{mkT}\frac{1}{\ln(10)} \ . \tag{3.42}$$

Therefore, the subthreshold swing SS is:

$$SS = \left\{\frac{d\left[\log_{10}(I_{DS})\right]}{dV_G}\right\}^{-1} = \ln(10)\ m\ \frac{kT}{q} \qquad [\mathrm{V/dec}] \ . \tag{3.43}$$

Also m can be expressed in terms of the capacitances with Eq. (2.8), so:

$$SS = \left\{\frac{d\left[\log_{10}(I_{DS})\right]}{dV_G}\right\}^{-1} = \ln(10)\ \frac{kT}{q}\left(1+\frac{C_D}{C_{ox}}\right) \qquad [\mathrm{V/dec}] \ . \tag{3.44}$$

4 Discussion and conclusions

After analyzing carrier transport in subthreshold region, we came to the conclusion that the subthreshold drain-source current is a diffusion current ultimately originated from thermionic emission. As a consequence, the amount of current flowing through the channel depends on the fraction of electrons able to thermionically surpass the potential barrier, whose height varies depending on the different voltage conditions of the device.

From the subthreshold drain-source current, an analytical expression for the subthreshold slope SS was obtained, making use of its definition as the inverse of the characteristic's slope:

$$SS = \left\{ \frac{d\left[\log_{10}(I_{DS})\right]}{dV_G} \right\}^{-1} = \ln(10)\ m\ \frac{kT}{q}. \tag{4.1}$$

The objective of obtaining an analytical expression was to see which factors the SS directly depends on, and therefore, to gain an insight on the nature of the 60 mV/dec limitation.

Let us thus proceed to analyze the different terms of Eq. (4.1).

- The logarithm term comes exclusively from the logarithmic scale used to represent the transfer characteristic with the objective of better observing the subthreshold region. This value is a fixed value and does not stem from any physical property.

- The kT/q factor is, by definition, the thermal voltage and gives the SS a direct dependence with temperature. This dependence has a physical reason behind. The subthreshold slope indicates how much voltage is needed to increase the current by a factor of 10, and if the current depends upon temperature, so will SS. As previously explained, the subthreshold current is a diffusion current originated by an unbalance of carrier concentration at the edges of the channel. Those carrier concentrations are given by the fraction of electrons able to surpass the potential barrier by thermionic emission. This fraction corresponds to the electrons located at energy levels above the barrier and it is derived from Fermi-Dirac distribution function and the density of allowed states. Therefore, it presents a direct dependence on temperature associated to the Fermi-Dirac term, yielding the current's thermal dependence. That way, the kT/q term is directly linked to the thermionic emission of carriers over the potential barrier of the channel.

  As a consequence of this dependence, the value of this term could be reduced by decreasing the temperature at which the circuit works. Nonetheless, this is not an optimal solution because cooling considerably increases the device cost [6]. Besides, cooling solutions are bulky and markedly bigger in size than the circuit. Therefore, they pose a problem for nowadays applications, where reduced device dimensions is one of the main objectives.

- The last term to analyze is the body effect coefficient m. As previously explained, this parameter tells us the fraction of the applied gate voltage that drops across the semiconductor. It has a direct dependence on the capacitances of the MOS structure:

  $$m = 1 + \frac{C_D}{C_{ox}} \ . \tag{4.2}$$

  The body effect coefficient m can take as a minimum value 1, corresponding to the case where all the gate voltage reaches the surface of the semiconductor. Certainly, this is the ideal limit yet a typical m value for a MOS transistor would be in between 1.1 and 1.3 [11].

  Despite, there are some ways to obtain a value of m close to 1. One would be to reduce the depletion capacitance C_D, that is, to increase the depletion region width W_D of the channel. This parameter depends inversely on the doping concentration, therefore by reducing it we will increase the depletion width. Light doping leads to a small amount of charge in the semiconductor, which produces a small electric field in the oxide and a correspondingly small voltage drop across it [11]. Therefore, almost all of the applied gate voltage will drop across the semiconductor leading to an m closer to 1. Nevertheless, this comes along with some drawbacks, because with less doping concentration, there will be less charge carriers to form the channel and therefore less current in the on-state.

  Another possible way to reduce m would be to increase the oxide capacitance C_{ox} reducing the thickness of the insulating layer t_{ox}. Nonetheless, the oxide thickness cannot be reduced much less than roughly a nanometer [17]. Beyond that, too much leakage current is generated due to quantum tunneling between the channel and the gate, resulting in high static power consumption. This leakage exceeds the requirements of some applications (e.g., DRAM) already at 2.5–3 nm. Either way, the minimum pure SiO_2 gate insulator thickness for high-performance applications is in the 1.0–1.5-nm range [7].

To obtain the limit of 60 mV/dec, numerical values for all of the terms in Eq. (4.1) have to be considered. The logarithm is ln(10) ≈ 2.3. The thermal voltage value will be chosen at room temperature, considering it the standard temperature at which devices work. Its value at $T = 300$ K is equal to 26 mV. Also, the optimal value of the body effect coefficient is $m = 1$. Therefore, substituting yields:

$$SS = \left\{ \frac{d\left[\log_{10}(I_{DS})\right]}{dV_G} \right\}^{-1} = 2.3 \cdot 1 \cdot 26 \simeq 59.87 \simeq 60 \text{ mV/dec} \ . \tag{4.3}$$

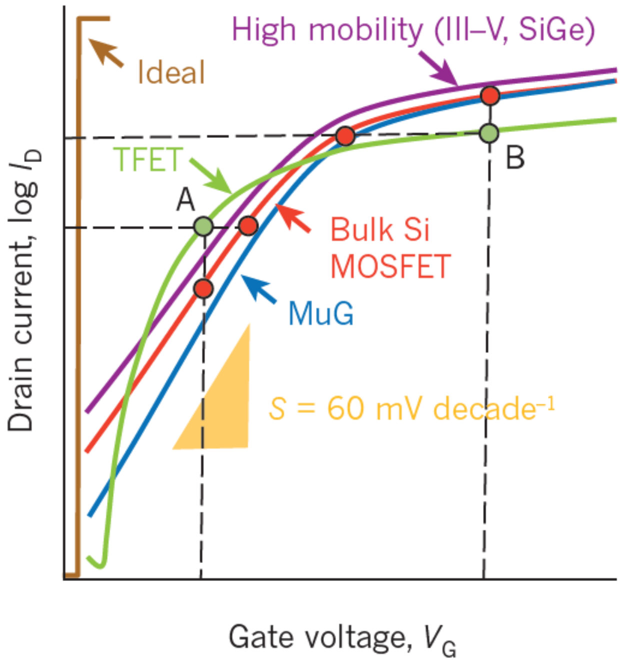


Figure 4.1: Transfer characteristics of different transistors [10].

The limiting term in the subthreshold slope is, as stated before, the thermal voltage. It restricts the amount of electrons available to conduct current. This thermionic limitation stems from the physics behind the functioning of the MOSFET, associated to semiconductors behaviour, and therefore it cannot be modified. That way, the 60 mV/dec subthreshold slope constitutes an infrangible limit for CMOS technologies. It should be noted that subthreshold swings below 100 mV/dec are considered satisfactory values for conventional MOSFETs [11], and, in effect, they seldom reach the 60 mV/dec bound.

However, there are some devices based on MOSFET's structure, with slight modifications, that approach the 60 mV/dec limit.

One of the most prominent ones is the Extremely Thin SOI MOSFET [11]. This fully depleted structure has a body effect coefficient $m = 1$ and it can thereby achieve the lowest possible SS [11]. Such transistors are beneficial for reducing the power consumption of the circuits, as it was previously discussed.

Another popular device is the FinFET, a quasiplanar double-gate device that effectively suppresses short channel effects and is not too distant from the conventional MOSFET in terms of fabrication [8]. The FinFET reaches values of SS close to 60 mV/dec, but it increases with scaling.

The problem with this type of devices is they that are also governed by thermionic emission processes over a potential barrier (see Fig. 4.2), and therefore have inevitably the same limitation as the bulk MOSFET.

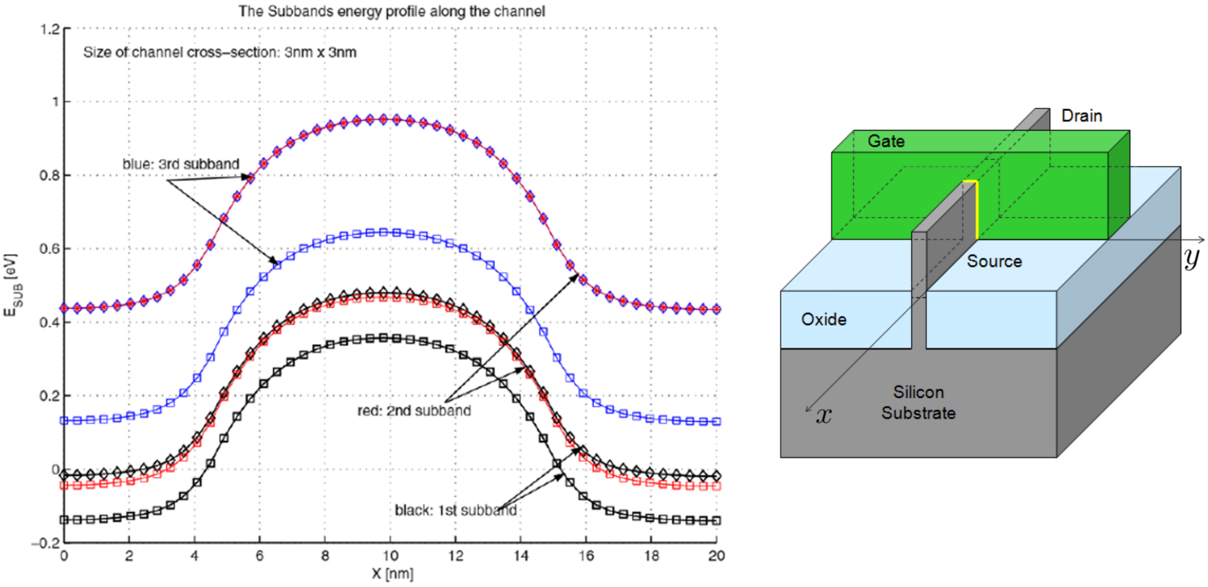


Figure 4.2: Structure and energy band diagram along the channel (x-direction) for a FinFET with 1-nm gate oxide and 10-nm gate length, with no drain or gate voltage applied. Modified figure from [18] and [9].

Consequently, to beat the thermionic limit and minimize the SS value beyond 60 mV/dec, emerging devices based on different physical principles have to come into play. One of the most promising ones at the state of the art is the Tunneling Field Effect Transistor (TFET). Tunnel FETs avoid the thermionic limit by using quantum-mechanical band-to-band tunnelling, rather than thermal injection, to inject charge carriers into the device channel. Although the current-control mechanism in the TFET is different, the device bears a strong resemblance to the MOS transistor. It has the same basic configuration of source, drain and gate and similar electrical behaviour when wired into circuits. Tunnel FETs, based on ultrathin semiconducting films or nanowires, could achieve a 100-fold power reduction over CMOS transistors, reaching subthreshold swings of 40 mV/dec [10].

Fig. 4.1 is a qualitative comparison of three engineering solutions to improve the characteristics of the bulk silicon MOSFET switch (red): a multigate device (MuG, blue) for improved electrostatics; a high-mobility channel (purple) using group III–V and SiGe materials; and a TFET (green). At operation point A, because of its subthermal subthreshold swing, the TFET offers not only an improved I_{on}/I_{off} ratio but also a superior performance and a power saving at the same performance as a MOSFET.

Therefore, as many authors defend [10] [13], today Tunneling Field-Effect Transistors represent the most promising steep-slope switch candidate.

The fundamental physical limitations that CMOS technologies face nowadays are hard to overcome, as we have seen throughout this work. This may bring an end to the predominance of MOS transistors in microelectronics. At the same time a promising line of research arises: transistors based on different physical properties that achieve lower subthreshold slopes than the MOSFET.

It is not clear yet which emerging device will substitute MOSFETs, or if any ever will. What is certainly known is that, in the near future, innovations will have to be made in directions other than those that have driven semiconductor technologies so far. The research community in this field must continue in the quest for a paradigm shift that surmounts the development problems we are facing these days, and eventually, to the next big technological revolution.

References

[1] Abbas, Zia, & Olivieri, Mauro. 2014. Impact of technology scaling on leakage power in nano-scale bulk CMOS digital standard cells. *Microelectronics Journal*, **45**(2), 179–195.

[2] Biljanovic, P., & Suligoj, T. 2000 (May). Thermionic emission process in carrier transport in pn homojunctions. *Pages 248–251 of: Proceedings of 10th Mediterranean Electrotechnical Conference. Information Technology and Electrotechnology for the Mediterranean Countries. Proceedings.*, vol. 1.

[3] Cressler, John D. 2009. *Silicon Earth: Introduction to Microelectronics and Nanotechnology.* Cambridge University Press.

[4] De, V., & Borkar, S. 1999 (Aug). Technology and design challenges for low power and high performance microprocessors. *Pages 163–168 of: Proceedings. 1999 International Symposium on Low Power Electronics and Design (Cat. No.99TH8477).*

[5] Dennard, Robert H, Gaensslen, Fritz H, Yu, Hwa-Nien, Rideout, V Leo, Bassous, Ernest, & Leblanc, Andre R. 1974. Design of ion-implanted MOSFET's with very small physical dimensions. *IEEE Journal of Solid-State Circuits*, **9**(5), 256–268.

[6] EECS Department, UC Berkeley. 2005. *Ch. 7 MOSFET Technology Scaling, Leakage Current, and Other Topics.* Lecture Notes from Course "Integrated Circuit Devices". https://inst.eecs.berkeley.edu/ ee130/sp06/chp7full.pdf.

[7] Frank, David J, Dennard, Robert H, Nowak, Edward, Solomon, Paul M, Taur, Yuan, & Wong, Hon-Sum Philip. 2001. Device scaling limits of Si MOSFETs and their application dependencies. *Proceedings of the IEEE*, **89**(3), 259–288.

[8] Hisamoto, D., Wen-Chin Lee, Kedzierski, J., Takeuchi, H., Asano, K., Kuo, C., Anderson, E., Tsu-Jae King, Bokor, J., & Chenming Hu. 2000. FinFET-a self-aligned double-gate MOSFET scalable to 20 nm. *IEEE Transactions on Electron Devices*, **47**(12), 2320–2325.

[9] Intel. *Innovations in 22 nm transistor technology improve performance and energy efficiency.* Accessed on June 2019: https://www.intel.com/content/www/us/en/silicon-innovations/revolutionary-22nm-transistor-technology-presentation.html.

[10] Ionescu, Adrian M, & Riel, Heike. 2011. Tunnel field-effect transistors as energy-efficient electronic switches. *Nature*, **479**(7373), 329.

[11] Lundstrom, Mark. 2017. *Fundamentals of Nanotransistors.* World Scientific.

[12] Moore, G.E. 1965. Cramming more components onto integrated circuits. *Electronics*, **38**(8), 114–117.

[13] Nikonov, Dmitri E, & Young, Ian A. 2013. Overview of beyond-CMOS devices and a uniform methodology for their benchmarking. *Proceedings of the IEEE*, **101**(12), 2498–2533.

[14] Pulfrey, David L. 2010. *Understanding modern transistors and diodes.* Cambridge University Press.

[15] Razavi, B. 2013. *Fundamentals of Microelectronics.* 2nd edn. Wiley Global Education.

[16] Sakurai, T. 2003 (Feb). Perspectives on power-aware electronics. *Pages 26–29 vol.1 of: 2003 IEEE International Solid-State Circuits Conference, 2003. Digest of Technical Papers. ISSCC.*

[17] Seabaugh, Alan. 2013. The tunneling transistor. *IEEE spectrum*, **50**(10), 35–62.

[18] Shao, Xue, & Yu, Zhiping. 2005. Nanoscale FinFET simulation: A quasi-3D quantum mechanical model using NEGF. *Solid-State Electronics*, **49**(8), 1435–1445.

[19] Stevenson, Richard. 2013. Changing the channel. *IEEE Spectrum*, **50**(7), 34–39.

[20] Streetman, Ben G, & Banerjee, Sanjay Kumar. 2016. *Solid State Electronic Devices: Global Edition.* Pearson Education.

[21] Sze, S. M. 1981. *Physics of semiconductor devices.* 2nd edn. New York: John Wiley and Sons.

[22] Taur, Yuan, & Ning, Tak H. 2009. *Fundamentals of Modern VLSI Devices.* 2nd edn. Cambridge University Press.

[23] Waldrop, M Mitchell. 2016. More than Moore. *Nature*, **530**(7589), 144–148.

[24] Wong, Ban, Mittal, Anurag, Cao, Yu, & Starr, Greg W. 2005. *Nano-CMOS circuit and physical design.* John Wiley & Sons.