

Neuromorphic Engineering 1

Lecture notes from Fall 2021

Authors and Contributors:

Lioba Schürmann (lschuermann@student.ethz.ch)
Yassine Taoudi-Benbekroun (ytaoudi@student.ethz.ch)
Wencan Huang
Pietro Bonazzi
Marc Steiner

<https://www.overleaf.com/project/61bdc6edc995887846caa1c9>
<https://github.com/yassinetb/Neuromorphic-Engineering-1-Lecture-Notes-2021>
<https://www.taskade.com/d/zhPqQZgDQQ37hGVA?share=view&view=U8rVUTCoq8SxT5q1>

December 2021

Disclaimer

These lecture notes were originally written to help the authors consolidate in a "stupid-proof" manner the content of the fascinating but very challenging Neuromorphic Engineering 1 (NE1) course. They are also meant to help other students of the current or following years to digest the content of the NE1 lectures, to prepare for the exam and, hopefully, better appreciate the subtleties of the field. They are almost exclusively based on the content taught during Fall 2021 by Prof. Shi-Chii Liu, Prof. Tobi Delbrück Prof. Giacomo Indiveri at ETHZ. **We do not claim originality nor completeness.** The containing information is thus freely gathered and often copied word-by-word from the lecture materials or the "Analog VLSI" textbook ¹ written by the previously mentioned professors and others (we generally refer to this as the "Textbook"). These lecture notes also reproduce content from other sources such as books, blogs, papers, videos etc... We tried to always cite our references when they did not come directly from the Textbook or Lecture Notes.

Organisation

The content is organised similarly to the Fall 2021 course schedule, thought it is not exactly the same. We chose to separate lecture 2 and 3 into two distinct chapters: Transistor Physics and Transistor Operation. This structure, similar to the Textbook, felt more natural and clearer. We also grouped both Photosensor and Photoreceptor lectures. The chapters are best read in order, as much of the content from a given chapter builds on the content from previous chapters. Each chapter contains the following:

- Short introduction about the objectives of the lecture, and some contextualization of the topic into the field of Neuromorphic Engineering and beyond.
- List of all topics previously covered that should be understood in order to properly digest the content of the chapter.
- Content of the chapter, organized, to the best of our capability, in a logical manner (sometimes not exactly matching the lecture structure).
- Lab Summary and takeaways.
- List of all concepts to know for the exam. These are often (though not exclusively) taken from the "Study Winter Sheet"².

Additionally, we added the following extra chapters for reference:

- Chapter 0: Stupid-Proof explanations of fundamental Electrical Engineering / Physics / Neuroscience for the uninitiated. Warning: This chapter may contain approximations and inaccuracies.
- Circuit recap - a revision sheet for the most important circuits, their function, equations and diagram.
- Advice for exam preparation.
- Important constants (It is being said that Tobi likes to ask about these in the exam!)

¹<https://mitpress.mit.edu/books/analog-vlsi>

²This is a revision sheet available on OLAT that should guide revision. However, one should note that many topics mentioned in this study sheet were not covered during Fall 2021, and that several important things that we believe should be known are not mentioned in the sheet. We also included it in the Appendix for reference.

Author Contribution

The project of writing these lecture notes emerged from discussions about the challenging NE1 course between Lioba Schürmann and Yassine Taoudi Benchekroun in early December 2021. Though our original ambition was to better prepare for the exam, it quickly became a project with bigger purpose: building a digestible guide to all present and future students of the challenging NE1 course, and Neuromorphic Engineering as a field. This shift of purpose is a result of the strong interest we feel towards this fascinating field of research - and it was particularly emphasized when Wencan Huang and Pietro Bonazzi joined us in early January, as they shared our enthusiasm for the project. Pietro and Wencan joining us truly gave another breath to this project.

We now all envision this as an *open source project* and very much look forward to adding contribution from students. We invite all readers to point us to the details we may have omitted, to the elements we may have described in inadequate fashion, or anything that could improve the overall quality of this humble work. It will be our honour to receive your contribution and suggestions and add you to our list of contributors.

The individual contribution of authors is as follows:

- Organization, coordination and structuring:

- Yassine Taoudi Benchekroun

- Chapter Authorship:

Some chapters were written by Yassine, others by Lioba. We are choosing not disclose which ones we wrote to avoid being *purposely* asked in the oral exam about chapters we did not write :)

- Lab takeaways:

- Wencan Huang: Odd-numbered labs.
 - Pietro Bonazzi: Even-numbered labs.

- Editing:

- All others contributed to editing and reviewing each other's part. Wencan Huang was particularly helpful in helping with the device physics and transistor operation sections, which were the most challenging to write of the whole document.

Contact

If you:

- Would like to contribute to the writing of these lecture notes,
- Spot an error or inaccuracy in the writing of these lecture notes,
- Have any questions, comments or complaints about the lecture notes,

Please feel free to raise an issue on the Github repository or to contact one of the authors directly at: ybenchekroun@outlook.com.

Contents

0 Fundamentals	8
0.1 Basics of Electronics	8
0.1.1 Fluid Model: The Key to Understanding Electronics	8
0.1.2 Charge	9
0.1.3 Electric Field	9
0.1.4 Voltage	10
0.1.5 Current	10
0.1.6 Resistance	11
0.1.7 Ohm's law	12
0.1.8 Capacitance	13
0.1.9 DC vs AC and exponential notation	14
0.1.10 Basics of Parallel and Series circuits	16
0.1.11 Kirchoff's Voltage and Current Laws	16
0.2 Basics of Computational Neuroscience	17
0.2.1 The Neuron	17
0.2.2 Electrical Perspective of the cell	18
0.2.3 The Action Potential	19
0.2.4 The synapse	20
0.2.5 Network and Computation	22
0.3 Basics of Machine Learning	23
0.3.1 The Perceptron	23
1 Neuromorphic Engineering: History, objectives and challenges	24
1.1 Modern Neuroscience	24
1.1.1 Cajal and Golgi: the birth of Neurons	24
1.1.2 Hodgkin and Huxley: The neural cell as a computational unit	25
1.1.3 Hubel and Wiesel Mountcastle: Experimental work on Neuron's Input to Output	25
1.1.4 Model of the brain as a complex system with detailed individual component behaviour	25
1.2 Modern Computing	25
1.2.1 The transistor	25
1.2.2 Perspective shifts and realizations: Von Neuman, Marvin Minsky, Feynmann and Carver Mead	25
1.3 Neuromorphic Engineering	25
1.3.1 Carver Mead and the Caltech Graduate Course	25
1.3.2 First Breakthroughs: Misha Mahowald and Silicon Retina	25
1.3.3 From research breakthroughs to deliverables	25
1.4 The rise of Machine Learning	25
1.5 Modern challenges and objectives	25
1.5.1 Energy considerations	25
1.5.2 Efficiency considerations	25
1.6 Challenges	25
1.7 List of complementary readings	25
2 The Essential Physics Behind the Transistor	27
2.1 Silicon: the magic semi-conductor	27
2.1.1 Prelude: Material Conductivity	27
2.1.2 Thermal Consideration	29
2.1.3 Silicon: structure, doping and properties	30
2.2 Understanding the PN junction and the Diode	33
2.3 MIS Capacitance Structure	36
2.4 Test yourself	41

3 Transistor Operation	42
3.1 Building an Intuition of the Transistor	42
3.1.1 Understanding the transistor idea with Hydraulic Analogy	42
3.2 MOSFET Structure and basic function	43
3.3 Understanding Sub-threshold Current using Boltzmann Distribution (Written by Wencan Huang)	46
3.4 Subthreshold Operation	49
3.4.1 Prelude: Drift and Diffusion Current	49
3.4.2 Let's start deriving equations	50
3.5 Superthreshold Operation	53
3.6 pFET MOSFET	56
3.7 Bulks, wells and biasing the MOSFET Bulks	57
3.8 Transistor Conductance	58
3.9 Second Order Effects	58
3.9.1 Prelude: How transistor width and length impact operation	58
3.9.2 Transistors in real life, and the problem of mismatch	59
3.9.3 The Early Effect	59
3.9.4 The Body Effect	62
3.9.5 Drain Induced Barrier Lowering (DIBL)	63
3.10 Impact Ionization	64
3.11 Concluding thoughts on Subthreshold Regime	65
3.12 Laboratory : Transistors above threshold	65
3.12.1 Voltage threshold and Beta	65
3.12.2 Early voltage	65
3.13 Test Yourself	66
4 Static Circuits	67
4.1 Single Transistor Circuits	67
4.1.1 The Current Source	67
4.1.2 Linear Resistor	68
4.1.3 Non Linear Current-Voltage / Voltage-Current Converter	69
4.1.4 Diode Connected Transistors	69
4.2 Two Transistor Circuits	70
4.2.1 Current Mirror	70
4.2.2 Intrinsic Voltage Gain	71
4.2.3 Source Follower	72
4.3 Three (and more) Transistor Circuits	73
4.3.1 The differential pair	73
4.3.2 The current correlator	75
4.3.3 The Bump-antibump circuit	76
4.4 Laboratory : Static Circuits	76
4.4.1 N-FET differential pair circuit (NDP)	77
4.4.2 Bumb antibump	77
4.5 Test yourself	78
5 The Transconductance Amplifier	79
5.1 Architecture	79
5.2 Transconductance Amplifier Function	80
5.2.1 Let's assume everything is in Saturation	80
5.2.2 Let's stop assuming that everything is in Saturation	81
5.2.3 Transconductance Amplifier as a Voltage Amplifier	83

6 Linear Systems Theory	85
6.1 Preliminary to Resistor Capacitor Circuits	85
6.1.1 Complex Exponentials	86
6.2 Step and Delta function	86
6.2.1 The Heaviside-Laplace Transform	86
6.2.2 Transfer Function	87
6.3 Resistor-Capacitor Circuits	87
6.3.1 Solving Low pass Integrator RC circuit	87
6.3.2 Solving High pass Differentiator CR circuit	89
6.3.3 Frequency Domain Analysis	89
6.3.4 Why are they called integrators and differentiators	91
6.3.5 Summary about filters	91
6.4 VLSI Integrators and Differentiators	92
6.4.1 Unity Gain Follower	92
6.4.2 Follower Integrator	93
6.4.3 Delay Lines	95
6.5 Laboratory : Integrator Circuits	96
6.5.1 Time-domain response of small signal	96
6.5.2 Time-domain response of large signal	97
6.5.3 Frequency-domain response	97
6.6 Things you should know	98
7 Current Mode and Winner Take All	99
7.1 Translinear Circuits	99
7.1.1 Short intro	99
7.1.2 Translinear principle	99
7.2 The Current Conveyor	100
7.2.1 Introduction	100
7.2.2 Basic subthreshold current conveyor	102
7.3 The current conveyor as a multiplier	103
7.4 The Gilbert Normalizer	103
7.5 Winner Take All circuit	104
7.5.1 $I_{in_1} = I_{in_2} = I_{in}$	105
7.5.2 $I_{in_1} \gg I_{in_2}$	105
7.5.3 $I_{in_1} = I_{in_2} \pm \delta I_{in}$	106
7.5.4 Experimental data	107
7.6 Things you should know	107
8 Silicon Synapses	108
8.1 VLSI Synapses in Pulse-based Neural Networks	108
8.2 Exponentially decaying integrator circuit	109
8.3 Log-domain Pulse Integrator	110
8.4 Diff-Pair Integrator (DPI) Synapse	111
8.4.1 Short-term Depression	113
8.5 Laboratory :Silicon Synapses	115
8.5.1 DPI synapse	115
8.6 Test Yourself	116
9 Silicon Neurons	118
9.1 Conductance-based silicon neuron	118
9.2 Axon-hillock circuit	119
9.3 Test Yourself	123

10 Photosensors and circuits	124
10.1 Prelude: Motivation	124
10.2 Light	124
10.2.1	125
10.2.2 Measuring light: Radiometry and photometry	125
10.3 Physics of Photosensors	126
10.3.1 Photoelectric Effect	126
10.3.3 Optical Absorption	126
10.3.4 Quantum Efficiency	127
10.4 Interlude: Human Vision and the Retina	128
10.4.1 General architecture of the visual system	128
10.4.2 The human eye	129
10.4.3 The retina	130
10.4.4 Human photoreceptors: Rods and Cones	130
10.4.5 Photo receptor structure and function	131
10.5 Silicon Photosensors	132
10.5.1 Common principles of Photosensors	132
10.5.2 Photoconductor	133
10.5.3 Photodiode	133
10.6 Operations of Photoreceptors	136
10.6.1 Dark Current	136
10.6.2 How to estimate incident light on the chip	136
10.6.3 Why is a log response desirable?	137
10.7 Logarithmic Photoreceptor	137
10.8 Reasoning through Source Follower Logarithmic Photoreceptor	137
10.8.1 Time domain response of source follower response	139
10.9 Reasoning through Transimpedance Logarithmic Photoreceptor	139
10.9.1 Time domain response of transimpedance logarithmic Photoreceptor	141
10.10 Laboratory : Photoreceptors	141
10.10.1 Static DC responses	141
10.10.2 Large signal transient response	142
10.10.3 Small signal transient response	142
10.11 Adaptive photoreceptor	145
10.12 Cascode and Miller Effect	147
10.13 Test Yourself	147
11 Complementary Notes and Topics	149
11.1 Revision Table	149
11.2 Know your constants - Tobi will ask you	149
11.3 Exam Tips	149
11.3.1 General comments on the exam format and expected level of details	149
11.3.2 Questions that they we think will be asked	149
11.3.3 How to practice your circuits	150
12 Appendix	151

0 Fundamentals

0.1 Basics of Electronics

In this section, we'll have a look at fundamental concepts in electronics. We'll also point you to the right references to understand the concept if our explanations are not enough. I believe the best way to grasp these challenging concepts is through water analogies: as we see water every day (as opposed to electrons), it is a lot easier to understand elementary fluid dynamics concepts than elementary electricity concepts. Fortunately, whoever created this world loved the concept of *Fractals* and self-similar patterns, so it's very easy³ to make accurate electricity analogies with fluids. Because this course goes in depth into electronics, it is assumed that all these concepts are already fairly familiar to the student. We have chosen to cover it just as a reminder or merely a pointer to what one should know before starting the course. If one is not fully comfortable with these concepts, time should *first* be spent understanding them before going further into the course. It is not needed to enter into deep theory, but one should be comfortable enough with the conceptual behaviour in order to understand the rest of the course.

0.1.1 Fluid Model: The Key to Understanding Electronics.

This section is largely inspired by Carver Mead's *Analog VLSI* textbook, which I cannot recommend enough.

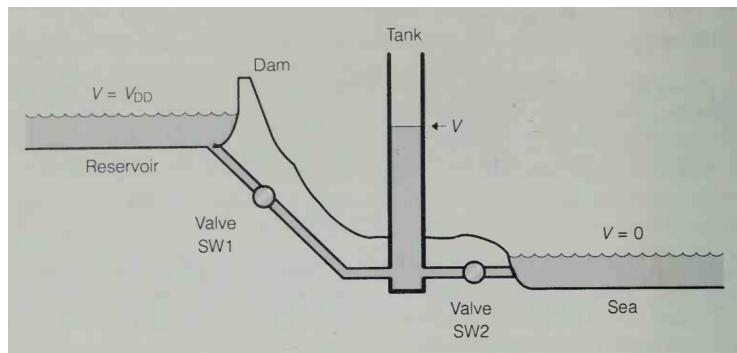


Figure 1: Hydraulic analogy of electronic (or neural) circuit. The power supply is a reservoir of water at potential V_{dd} . The reference potential is sea level, corresponding to *ground* of electrical circuits. The water level in the tank corresponds to the output voltage. The tank can be filled by opening the valve SW1 or emptied by opening valve SW2 - the valves can be compared to switches in electrical circuit. Adapted from Carver Mead *Analog VLSI* textbook

We are able to use fluid analogies to describe electrical dynamics because the fundamental component of electrical circuit (the charged particle - be it ions or electrons) is present in sufficient number that they cannot be accounted for individually - similarly to fluid dynamics where we cannot model at the scale of the individual fluid molecule. The quantity of water in figure 1 is analogous to the quantity of electrical charge Q . There is an underlying granularity to these quantities: water is made up of water molecules, charge in a neuron is made up of ions (more on that later) and charge in a transistor is made up of electrons. In all cases, we can discuss the quantity in terms of either the number of elementary particles or a more convenient macroscopic unit.

In electronics, the flow of the fundamental particle is called the **electrical current**. When speaking of fluids dynamics, gravity is a critical force that will impact much of the behaviour of this fluid - in electronics, the analogous gravitational "potential" is called **voltage**. In figure 1, it corresponds to the height at which the water stands compared to the reference point: sea level. It is quite straightforward to understand from this point of view that, exactly as the water

³Honestly it's almost scary how similar it is, you'd believe it's all the same thing. This makes me think of a slightly unrelated essay which thinks about a somewhat related epistemological problem: https://en.wikipedia.org/wiki/The_Unreasonable_Effectiveness_of_Mathematics_in_the_Natural_Sciences

which is up in the reservoir is drawn to the ground (but doesn't necessarily reach if the valves are closed), electrical charge at a given potential is attracted to the electrical ground (but doesn't necessarily reach if switches are "opened" i.e. not allowing electrons to flow). This is what potential difference/voltage represents. This leads to the following observation, which makes up for most of what needs to be understood about fundamental electricity behaviour: a *potential difference* in charge drives a *current flow* from a higher potential point to a lower potential point, and it flows relative to the extent by which we let it flow: how *conductive* (or resistive) the path between the high and low potential points is. This is what makes up for Ohm's law, where Current = Potential Difference / Resistance - we'll get to this again later. Resistance/Conductivity in this analogy corresponds to the diameter and material used in the pipes linking the reservoir to the sea - the larger the diameter and the smoother the material of the piper, the easier water can flow, the opposite also holds.

It is interesting to note that we can measure the height of water from any reference point we choose. There is however one particular advantage to choosing "sea level" as the reference potential: the height will always be positive or zero, thereby making calculations easier. In electrical circuits, this point of reference is called ground, and it is at 0 Volts.

You might wonder what happens when we run out of water in the reservoir. In fluid dynamics we'd need a pump (and thus some energy) to bring back water to the top. This is somewhat similar in electrical systems, which typically use an electrochemical process (such as in batteries) to keep on bringing electrical charge into the "reservoir".⁴

Now that this has been introduced, we can look in a bit more details into each of the concepts that should be understood, and keep on going back to water analogies to make things clear. Let's start with the fundamental electrical component: electrical charge.

0.1.2 Charge

This is the water molecule in the hydraulic analogy: the fundamental component which behaviour we describe. In fact, we never really describe the behaviour of the individual component, but rather the behaviour of the aggregate of individual components. Electric charge is the physical property of matter that causes it to experience a force when placed in an electromagnetic field. Think of a charge in an electromagnetic field as you think of some matter in a gravitational field - it is subject to forces because it has a mass. Electrical charge is subject to a force because it is not "neutral". Electric charge can be positive or negative (commonly carried by protons and electrons respectively). A somewhat fundamental electrical charge can also be an ion rather than electron or proton. An ion is an atom which has lost or gained one electron through some chemical process - as it has lost a charged particle, it therefore becomes charged itself, either positively or negatively. Alike charges repel each other and unlike charges attract each other. An object with an absence of net charge is referred to as neutral. We define charge with the symbol Q and with the units of *Coulombs*. An electron has, by convention, a charge of $1.6 \cdot 10^{-19}$ Coulombs.

0.1.3 Electric Field

An electric field is the physical field that surrounds electrically-charged particles and exerts force on all other charged particles in the field, either attracting or repelling them. By convention, the electric field vector points from positive to negative. Electric field also refers to the physical field for a system of charged particles. It's conceptually a bit similar to Newton's gravitational law, where all matter possessing a mass exerts a force on other masses, depending on the distance that separates them (and the Gravitational Constant). Understanding electric field is important to understand the physics behind transistors, and it actually is not so obvious to grasp. It would take me quite some time to explain it well, as it is not a concept that the water analogy can explain easily - I thus recommend having a look at the *Khan Academy*⁵ explanation as I won't

⁴There are obviously many other processes to keep the level of charge in the reservoir to a certain level, but no need to get into that. Just know that it doesn't get there magically, energy is needed to get it there. This is why batteries run out, or why you don't have electricity in your house if you don't pay your electricity bill - in both cases, the supply of charge in the reservoir stops.

⁵<https://www.khanacademy.org/science/hs-physics/x215e29cb31244fa1:types-of-interactions/x215e29cb31244fa1:electric-and-magnetic-fields/v/electric-field-definition>

be able to do anything better than this.

0.1.4 Voltage

Voltage⁶, (V) is the potential energy of an electrical supply stored in the form of an electrical charge. Voltage can be thought of as the force that pushes electrons through a conductor and the greater the voltage the greater is its ability to “push” the electrons through a given circuit. The difference in voltage between any two points, connections or junctions (called nodes) in a circuit is known as the Potential Difference, (p.d.) also commonly called the Voltage Drop. The Potential difference between two points is measured in Volts with the circuit symbol V , or lowercase “v“.

Voltage is always measured as the difference between any two points in a circuit and the voltage between these two points is generally referred to as the “Voltage drop“. Note that voltage can exist across a circuit without current, but current cannot exist without voltage and as such any voltage source whether. This makes sense in the hydraulic analogy - there can be a height difference but no water flowing but there cannot be a flow without height difference⁷.

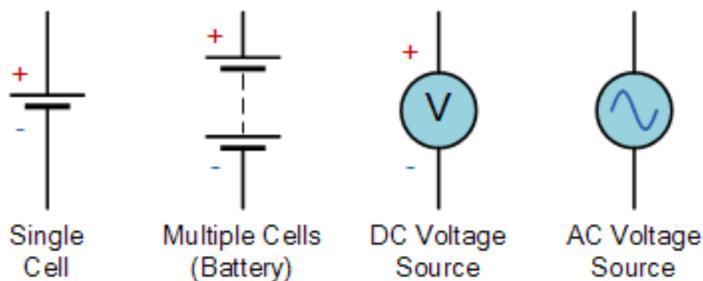


Figure 2: Voltage symbol. Adapted from https://www.electronics-tutorials.ws/dccircuits/dcp_1.html

0.1.5 Current

Electrical Current⁸ (I), is the movement or flow of electrical charge (most usually electrons) and is measured in Amperes, symbol I (or i), for "intensity". It is the continuous and uniform flow (called a drift) of charge around a circuit that are being “pushed” by the voltage source (height difference). In reality, electrons flow from the negative terminal to the positive terminal of the supply (as electrons are negative and thus attracted by positive point); however, for ease of circuit understanding conventional current flow assumes that the current flows from the positive to the negative terminal. This is why, by convention, the flow of current in circuit diagrams is represented by an arrow pointing from the positive to the negative node, despite electrons flowing from negative to positive node. This convention goes back to Benjamin Franklin: *“As far as the history goes, Ben Franklin imagined electricity as a type of invisible fluid that could build up or be absent from a material, or at least certain materials. He believed that when this invisible fluid built up the object was positively charged. When there was an absence of this fluid he called that material negatively charged. It turns out he got the concept right but the nomenclature backwards.”*. So it really is because of him that conventional current flow is not in the same direction as electron flow. Once you get used to it, it’s fine, but this can be quite annoying sometimes when you’re trying to visualize how circuits work.

⁶Section mostly copy-pasted from: https://www.electronics-tutorials.ws/dccircuits/dcp_1.html

⁷Yes, yes, I know, you can have water flowing in a flat surface but this is either because you push it somehow or because there is height drop somewhere further - in any case, water always flows as a result of some force being applied, either gravity or something human generated.

⁸Section mostly copy-pasted from: https://www.electronics-tutorials.ws/dccircuits/dcp_1.html

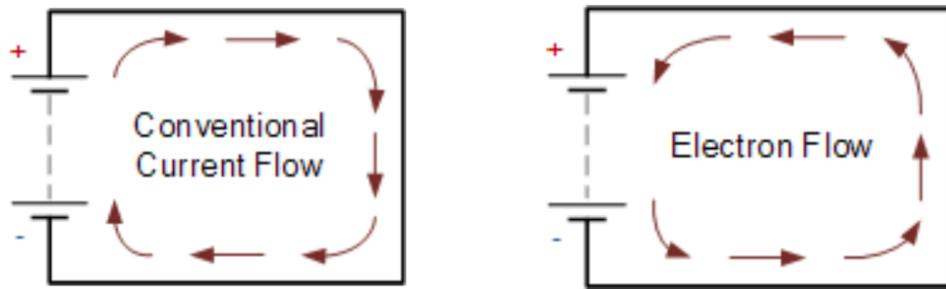


Figure 3: Electron flow and conventional current. Adapted from https://www.electronics-tutorials.ws/dccircuits/dcp_1.html

Current is defined by the following important relation:

$$I = \frac{dQ}{dt} \quad (1)$$

This becomes very clear in the hydraulic analogy, as we quantify the flow of water by the quantity of water flowing at a given point in time.

0.1.6 Resistance

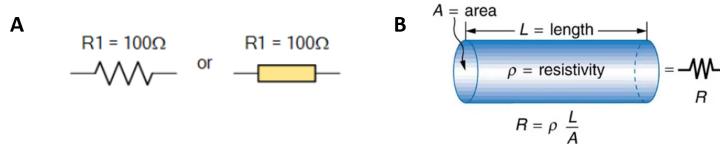


Figure 4: Resistor electric symbols. Adapted from https://www.electronics-tutorials.ws/resistor/res_1.html

In the hydraulic analogy figure 1, if both valves ⁹ are opened, water will flow from the reservoir into the sea at a finite rate, restricted by the diameter of the pipe through which it must flow. If the water level in the reservoir is increased, water will flow more quickly (up to a certain level), and the inverse is also true. The property of the pipe diameter and material, which restrict the flow of water, is called resistance. The electrical element possessing this quality is called a resistor.

The principal job of a resistor within an electrical circuit is thus to “resist” (hence the name Resistor), regulate or to set the flow of electrons (current) through them by using the type of conductive material from which they are composed. Resistors can also be connected together in various series and parallel combinations to form resistor networks which can act as voltage droppers, voltage dividers or current limiters within a circuit. This is analogous to separating an original pipe into multiple pipes where flow divides.

It is important to note that current is not affected by resistance in a *closed loop* circuit, because there is always **the same quantity of charge flowing in the loop**. In other words, current is constant everywhere in the closed loop circuit (admitting we do not add new elements), as long as there is current flowing. Voltage however, drops when encountering the different valves, simply because the height difference is reduced throughout water descent. This can seem a bit counterintuitive, but it will become clearer as we go through series and parallel circuits later.

Resistors are represented in electrical circuits as shown in figure 4. A and resistance is measured in *Ohms* (Ω). They are called “Passive Devices”, because they contain no source of power or amplification but only attenuate or reduce the voltage or current signal passing through them. This attenuation results in electrical energy being lost in the form of heat as the resistor resists the flow of electrons through it.

⁹keep in mind from the first section that valves are simply switches, allowing water (or current) to flow (or not). They can be thought of elements with infinite resistance when blocking flow, and 0 resistance when allowing flow

Most types of resistor are linear devices that produce a *voltage drop* across themselves when an electrical current flows through them. Different values of resistance produces different values of current or voltage. This can be very useful in Electronic circuits by controlling or reducing either the current flow or voltage produced across them we can produce a voltage-to-current and current-to-voltage converter.

The resistance of any substance depends on the following factors: 1) the length of the device (L), 2) the cross sectional area of the device (A), 3) the nature of material of the device, which has an inherent *resistivity* (ρ , measured in $\Omega \cdot m$), 4) the temperature of the device (T) (see Figure 4.B). All these variables also apply to evaluate pipes resistance in the water analogy!

$$R = \rho \frac{L}{A} \quad (2)$$

We say that a physical element or device has **infinite resistance** when current doesn't flow through it: it is an **insulator**. In practice, there is always a tiny tiny bit of current flowing, simply because infinite resistance is not realistic. To understand properly why this is the case, we would need to look into the physics of conduction, which will be covered in chapter 2 when introducing the transistor physics. For now, just imagine a big big wall - as big and strong as it is, if you have a strong enough water pressure applied on it, it may eventually break and allow current to pass. Another thing is that, even if it doesn't break, the wall cannot prevent water molecules which evaporate to pass to the other side. Moral of the story, even with very high resistance, you always have some flow, and there is no such thing as absolutely infinite resistance - it can always break. And exactly like in the water analogy, resistors can break if we apply too strong of an energy on them.

One last thing, it is often convenient to view an electrical circuit element in terms of its *willingness* to carry current rather than its reluctance to do so. As such, we often speak of *conductance* rather than resistance. Both properties are analogous and represent essentially the same idea. Conductance, represented with the letter G (sometimes g) is simply the inverse of resistance: $G = \frac{1}{R}$

0.1.7 Ohm's law

Simply, there is a linear relationship that relates Voltage and Current:

$$I = \frac{V}{R} = V \cdot G \quad (3)$$

A voltage of 1 Volt across a resistance of 1Ω will cause a current flow of 1 amp .

0.1.8 Capacitance

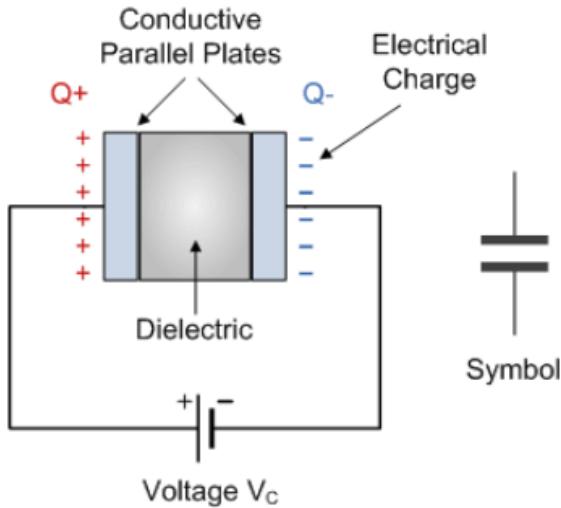


Figure 5: Capacitor structure and electric symbol. Adapted from <https://www.electronics-tutorials.ws/capacitor/cap1.html>

In the hydraulic analogy, the capacitor simply corresponds to the tank (with constant cross sectional area). Depending on cross sectional area, it will be able to hold a certain quantity of charge. It also needs time to charge up, and time to discharge. This leads to a shift in circuit analysis, where "steady state" analysis is not enough anymore - one must consider the charging and discharging properties of the capacitor as one would consider the flow relative to the tank filling up and emptying itself.

In its basic form, an electrical capacitor consists of two or more parallel conductive (metal) plates which are not connected or touching each other, but are electrically separated either by air or by some form of a good insulating material such as waxed paper, mica, ceramic, plastic or some form of a liquid gel as used in electrolytic capacitors. The insulating layer between a capacitor's plates is commonly called the Dielectric.

Due to this insulating layer, current does not flow through the capacitor as it blocks it allowing instead a voltage to be present across the plates in the form of an electrical charge.

Also, because capacitors store the energy of the electrons in the form of an electrical charge on the plates the larger the plates and/or smaller their separation the greater will be the charge that the capacitor holds for any given voltage across its plates. In other words, larger plates, smaller distance, more capacitance. This yields the following relation:

$$C = \frac{\epsilon_0 A}{d} \quad (4)$$

with C capacitance in Farads (F), ϵ permittivity of the dielectric ¹⁰, A area of the plate overlap in m^2 and d distance between the plates in meter.

The capacitor is thus a component which has the ability or "capacity" to store energy in the form of an electrical charge producing a potential difference (Static Voltage) across its plates, much like a small rechargeable battery. Capacitance is the electrical property of a capacitor and is the measure of a capacitor's ability to store an electrical charge onto its two plates with the unit of capacitance being the Farad (abbreviated to F) named after the British physicist Michael Faraday

By applying a voltage to a capacitor and measuring the charge on the plates, the ratio of the charge Q (in Coulombs) to the voltage V (in Volts) will give the capacitance value of the capacitor

¹⁰Permittivity is a measure of the electric *polarizability* of a dielectric. A material with high permittivity polarizes more in response to an applied electric field than a material with low permittivity, thereby storing more energy in the material

and is therefore given by the following important relation:

$$C = \frac{Q}{V}$$

One will then notice that $1\text{Farad} = 1\text{Coulomb per Volt}$

Although the charge is stored on the plates of a capacitor, it is more exact to say that the energy within the charge is stored in an “electrostatic field” between the two plates. When an electric current flows into the capacitor, it charges up, so the electrostatic field becomes much stronger as it stores more energy between the plates. Likewise, as the current flowing out of the capacitor, discharging it, the potential difference between the two plates decreases and the electrostatic field decreases as the energy moves out of the plates.

Still not sure you understand? If you feel like you still need to grasp the fundamental idea behind a capacitor, go watch the brilliant video by *The Engineering Mindset* on capacitors ¹¹.

0.1.9 DC vs AC and exponential notation

The explanations dealt with so far have constant voltage sources and are just taken at ”steady state” - think constant flow due to constant supply of water in the reservoir. Direct current (DC) is the flow of electric charge in a constant fashion. It is the steady state of a constant-voltage circuit. Most well-known applications, however, use a time-varying voltage source, yielding a time varying current, and time varying signals in general. This is relevant here not because we consider AC per se but rather because AC is a generalization of DC, and in some cases we need the time component (specifically with circuits with some capacitance). It is thus useful to be familiar with this concept, and most importantly, its notation. Alternating current (AC) is the flow of electric charge that periodically reverses direction. If the source varies periodically, particularly sinusoidally, the circuit is known as an alternating current circuit. Though batteries or power supplies are mostly used to produce a steady D.C. voltage source, A.C. (alternating current) voltage sources are used for domestic house and industrial power. This is for efficiency and transmission purposes, that we will not get into.

Figure 1 shows graphs of voltage and current versus time for typical DC and AC power. The AC voltages and frequencies commonly used in homes and businesses vary around the world.

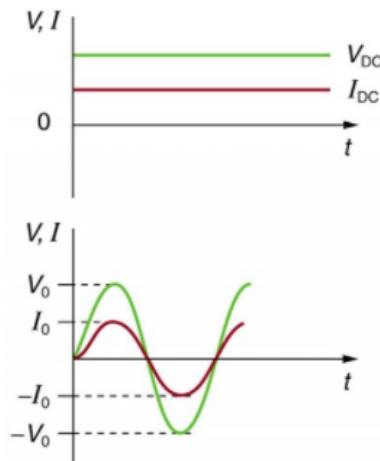


Figure 6: Alternating and direct current. Adapted from <https://courses.lumenlearning.com/physics/chapter/20-5-alternating-current-versus-direct-current/>

If we’re dealing with a different type of signal, we must use a different type of notation and consider the differences with DC. As can be seen on Figure 6, AC behaves in waves, so it is appropriate to treat it as such and use relevant notation that describes it properly. Also, we will

¹¹<https://www.youtube.com/watch?v=X4EUwTwZ110>

be using conventional notation for time varying voltage $u(t)$ and current $i(t)$ (as opposed to V and I that were previously used). We can now define ¹²:

$$u(t) = U \cos(\omega t + \phi) = U e^{j\omega t + \phi} \text{¹³} \quad (5)$$

Where U is the amplitude (peak value) of voltage, ω is the angular frequency, ϕ is the phase offset (more on the phase in a second).

Impedance There is a generalization of resistance, called *impedance* (represented by the symbol Z). It is also measured in Ohms. It is not really useful when working with DC current, but takes meaning when working with AC current, which has a complex component to it (which we call reactance but we don't care about that). **Ohm's law still applies in AC** and impedance is simply the ratio of the complex wave representation of sinusoidal voltage ($u(t)$) between its terminals to the complex wave representation of the current ($i(t)$) flowing through it. Impedance therefore possesses both magnitude and phase; resistance, on the other hand, has magnitude but no phase. This is why we say that impedance is a generalization of resistance, as resistance is only a special case of impedance. For impedance, we note:

$$Z = R + jX \quad (6)$$

where R is the real part resistance and the imaginary part X is the reactance. The magnitude of Z is $|Z| = \sqrt{R^2 + X^2}$, while the phase is $\phi = \arctan(\frac{X}{R})$. In general, ϕ is the phase difference between alternating voltage and current: $\phi = \phi_v - \phi_i$

Capacitance in complex analysis Capacitance is most often studied with complex analysis, so let's do that here and apply it to Ohm's law. While the current flowing through a resistor is given by the voltage applied to it, the current that flows through a capacitor is proportional to the voltage change.

$$i(t) = \frac{dQ}{dt} = \frac{d(CV)}{dt} = C \frac{du(t)}{dt} \quad (7)$$

Here, $i(t)$ is the current through time, which we defined previously as being the derivative of charge Q with respect to time. We also previously defined Q as being equal to the capacitance times voltage V (following from $C = Q/V$).

Differentiation can be conveniently performed in complex notation, yielding the following:

$$\frac{du(t)}{dt} = \frac{d(Ue^{j\omega t})}{dt} = j\omega U e^{j\omega t} = j\omega u(t) \quad (8)$$

Therefore, for a capacitor, $i(t) = C \frac{du(t)}{dt} = j\omega \cdot C \cdot u(t)$

¹²I am here assuming that you are familiar with exponential notation of periodic signals. If you aren't I suggest you spend some time reviewing this.

¹³This is derived from Euler's formula: $e^{jx} = \cos(x) + j\sin(x)$

0.1.10 Basics of Parallel and Series circuits

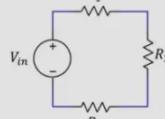
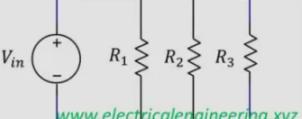
	Series	Parallel
How it looks		
Voltage	$V_{in} = V_1 + V_2 + V_3$	$V_{in} = V_1 = V_2 = V_3$
Current	$I_{series} = I_1 = I_2 = I_3$	$I_{in} = I_1 + I_2 + I_3$
Resistance	$R_{eq} = R_1 + R_2 + R_3$	$\frac{1}{R_{eq}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}$

Figure 7: Parallel and Series circuit comparison. Adapted from <https://www.electricalengineering.xyz/questions/top-5-differences-between-series-and-parallel-circuits/>

When dealing with electronic circuits, you can connect things in a variety of different ways, and some key things are to remember. Figure 7 does a great job at explaining this. Voltage drops across devices connected in series, but it is conserved across different branches of a circuit. Imagine having one branch separating into different branches on the same surface in the circuit: as all the pipes are at the same height from ground, they have the same voltage. Current is just the opposite, which make sense when you think of hydraulic analogy: it divides when separated into different branches and is maintained when flowing around a single loop. Resistance adds itself in series, which means that connecting two resistors together in the same branch makes the overall resistance stronger. This doesn't apply in parallel, where another relation need to be applied to find the total resistance of the circuit. If you have two branches and one has higher resistance than the other (pipe with small diameter), more current will flow through the lower resistance ones than the higher resistance one!

Capacitance, which is not shown on the figure, is actually exactly the opposite of resistance, it adds up in parallel and follows the same principle as resistance in parallel when it is connected in series.

0.1.11 Kirchoff's Voltage and Current Laws

One very important law is needed to understand many circuits we'll be studying: Kirchoff's voltage and current laws.

- Current law: All current flowing into a node (or junction) must be equal to the current flowing out of it. This means that there is no charge (think water) disappearing.

$$\sum I_{in} = \sum I_{out} \quad (9)$$

- Voltage law: In any complete loop within a circuit, the sum of all voltages across components which supply electrical energy (such as cells or generators) must equal the sum of all voltages across the other components in the same loop. This law is a consequence of both charge conservation and the conservation of energy. Think of a closed loop of reservoir at different height points which yield a continuous flow. Sum of heights will add up to 0.

$$\sum V_{total} = 0 \quad (10)$$

The following figure¹⁴ makes this much easier to understand.

¹⁴Adapted from: <https://www.sciencefacts.net/kirchhoffs-law.html>

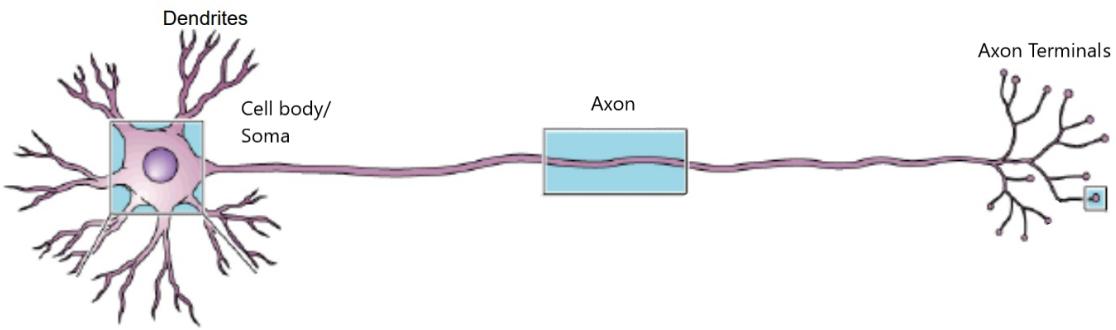


Figure 9: Schematic of a neuron. Adapted from the lecture notes of ETH Course: "Introduction to Neuroinformatics".

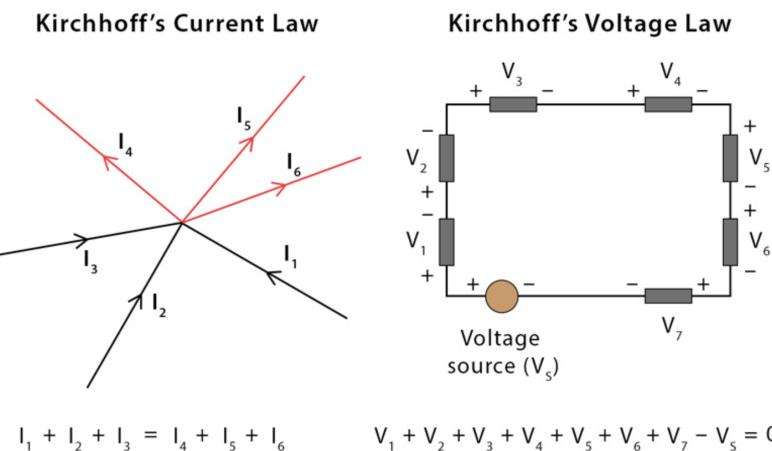


Figure 8: Kirchoff's current and voltage laws. Adapted from https://www.electronics-tutorials.ws/capacitor/cap_1.html

These are, I believe, all the necessary pre-requisites to understand the content of the module. Now we can build up from this and start understanding basics of transistors. But before heading there, let's look at some biology pre-requisites and try to explain them!

0.2 Basics of Computational Neuroscience

0.2.1 The Neuron

The most important component of our nervous system are neurons. They are able to encode sensory input from our environment into neural representations, process the retrieved information and decode it back into behavioural output, such as movements. Neurons are the fundamental element underlying the intelligent behaviour of virtually every animal, allowing them to interact with their environment, optimize survival etc... It is widely accepted that the number of neurons within a brain correlates with a specie's intelligence, somewhat similarly to a computer which becomes more powerful with higher number of *transistors*.. Figure 9 shows the schematic of a typical neuron.

There are 4 main components to the neuron: the **soma**, the **dendrites**, the **axon** and the **synapses** (not shown on figure, but we're getting there). These elements are enabling the individual neuron to connect to neighbouring neurons from which they can receive information, through synapses and dendrites, process this information, in the Soma and Synapses and propagate back information to other neurons, through axons. Intelligent behaviour is made possible by setting the

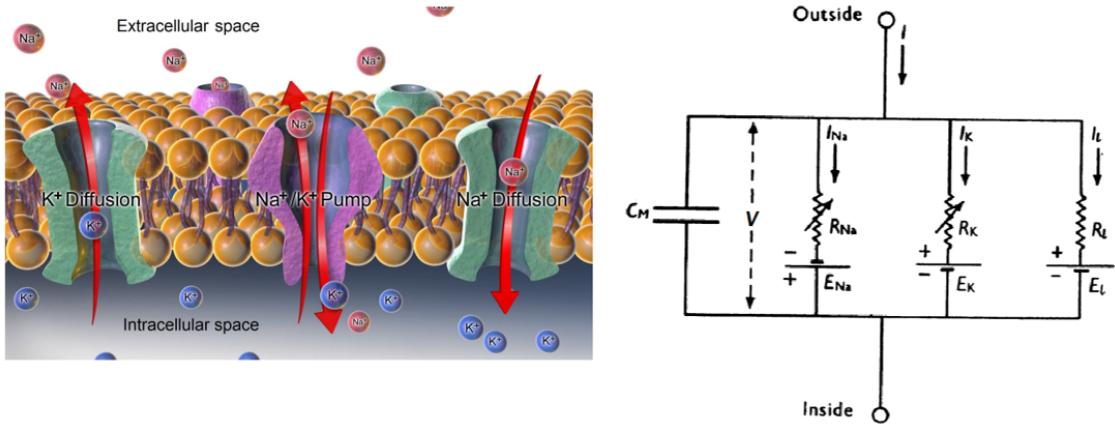


Figure 10: A) Biological cell lipid membrane, with the ion channels represented. Conductance based model of the biological cell. Adapted from the lecture notes of ETH Course: "Introduction to Neuroinformatics".

connections between different neurons and the "decision" rule to send a message accordingly. There exists many different subtypes of neurons, though they all exhibit relatively similar architecture and function.

Much of Neuroscience's research today is to understand the precise mechanisms by which neurons communicate and wire with each other. Though there are many perspectives one could focus one in order to study these dynamics, one particular paradigm of research has emerged in the recent years: **Computational Neuroscience**. This subbranch of Neuroscience devotes great effort to create *computational models* of brain function by modelling, often with complex mathematical equations, the behaviour of specific synaptic processes, individual neuron behaviour, and the dynamics of network of neurons. This has mainly been enabled by the rise of computing power and the improvement of imaging/recording technologies, which allowed observation, analysis and modelling of the fine dynamics exhibited in our brains. Though one particular scientific breakthrough gave birth this computational perspective: Hodgkin and Huxley's description of the **Conductance-Based Model** and the **Action Potential**.

0.2.2 Electrical Perspective of the cell

Hodgkin Huxley's conductance-based model ¹⁵ is essentially an electrochemical model of the biological cell (the neuron is a special biological cell).

As you know from high school biology, a cell (and also a neuron) has an intracellular body and a membrane which separates it to the extracellular space. The cell membrane consists of a double lipid layer that separates ions in the extracellular space from ions and charged proteins in the cytoplasm - it somewhat acts like an insulator. It so happens that both the intracellular space and extra cellular space are full of *ions*, mainly *potassium*, *calcium* and *sodium* ions. These ions are charged particles, and generate some electrical behaviour that we can analyze through the lense of electronic circuits. We can thus also speak of membrane voltage V_m , as the potential difference between the charged ions inside and outside the cell. We should also note that the cell can be seen as a **capacitor**, as it contains charged particles on both ends of an insulating layer. The insulating layer, actually contains specific **ion channels** that allow for specific ions to enter or exit the cell at specific moments. Therefore, to change the membrane voltage, it is necessary to charge the capacitance. The applied charge (Q) divided by the membrane capacitance (C_m) gives the membrane voltage (V_m): $V_m = Q/C_m$. We can see that for a given amount of applied charge, the smaller the membrane capacitance, the larger the membrane voltage change.

You can clearly see this on figure 10.A. Think of it like a dance club bouncer, allowing selected guests only to enter (or exit). Bouncer are less selective on Tuesday nights than Saturday nights:

¹⁵I recommend this excellent blogpost explaining the dynamics in more details: <https://www.scientifica.uk.com/learning-zone/understanding-the-cell-as-an-electrical-circuit>

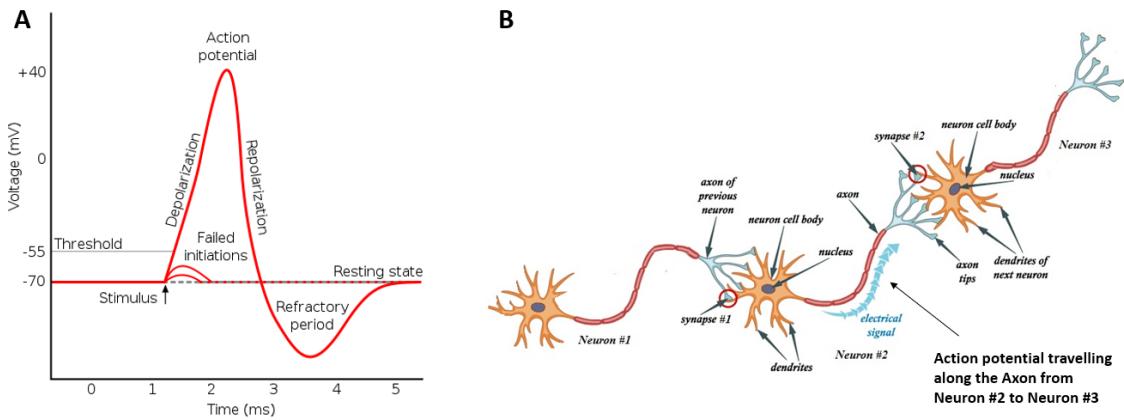


Figure 11: A) Action Potential Voltage change through time. B) Typical course of an action potential.

their resistance or **conductance** changes depending on the context. These basic structural components of the cell (and neuron) underlie the conductance based model that you can see on 10.B. We can see from the figure a few key components:

- Membrane Capacitance C_m : This is the capacitance that the insulating membrane carries by separating between the intracellular and extracellular space.
- The intracellular and extracellular space. Flow of ions (yielding current) must go through the different channels with specific conductance levels in order to enter or exit the cell.
- At rest (resting potential) the Sodium ions are in greater concentration outside the cell, and potassium ions inside the cell. This is represented with the potential levels E_{NA} and E_K configuration.
- A leak current I_L as the membrane is not a perfect insulator.

All biological cells communicate with each other via electrochemical signals, which means flow of ions entering and exiting the cells in precise fashion. This is very convenient because we can apply much of our existing electronics mathematical modeling to describe these behaviours! Now that we understand and have the intuition behind the language that cells communicate with, we should understand the principal messaging characteristic: Action Potential

0.2.3 The Action Potential

The neuron in its neutral state, i.e. when it's not receiving or sending any messages, has a negative potential compared to its surroundings. This potential is called the **resting potential** and it's about -70 mV. It is caused by different concentrations of ions in the intracellular and the extracellular space. The "messages" a neuron receives are sudden influx or ions from neighbouring cells. These are first processed at the synapse level and eventually transmitted to the Soma through the dendrites. These can either increase the neuron's resting potential, in which case they are *excitatory* input signals, or further decrease it and are *inhibitory* inputs. When a neuron's potential is increased so much that it becomes larger than its a specific **threshold voltage**, it sends out a message along its axon. This message is called an AP (AP) and its typical course is visualized in figure 11.

As can be seen from figure 11.A, once a neuron's potential crosses its threshold value, it experiences a large and fast increase of potential up to +40 mV compared to the extracellular space. This phase is called **depolarization** as the potential difference between the cell's intra- and extracellular space initially decreases. In the second phase, called the **repolarization**, the neuron's potential quickly decreases again. However, it doesn't stop at its initial resting potential but becomes even more negatively charged. This is called the hyperpolarization phase and it is denoted

as refractory period in the figure. The **refractory period** is the time it takes for the neuron's potential to return back to its initial state and during this time it is not possible, or at least a lot more difficult, to generate another action potential. The action potential is commonly called a *spike* and a neuron is said to fire when it generates a spike. The threshold voltage of a neuron is usually around -55 mV.

The idea of threshold voltage makes sense on a practical perspective: as there are *many* neurons, which constantly receive inputs from everywhere, either through noise or irrelevant signaling. You want neurons to select only the strong input messages, the ones that are somewhat unusual - you want neurons to filter out the noise. Functioning with a threshold voltage is a very clever way biology figured out to implement filtering! Of course, there are very complex dynamics behind this, and the way this precisely encodes information at the scale of the network is still an active domain of research. Though a critical part of this was understood through the research of Hubel and Wiesel and Donald Hebb, which we'll briefly touch upon in the next chapter. Essentially, you should remember two things: 1) Neurons that fire together wire together; 2) Specific neurons fire A LOT when they receive specific inputs they're sensitive to, and barely fire when these inputs are not received.

On another note, the AP can, in some sense, be considered a "digital" element¹⁶ as it either *fires* when voltage goes past the threshold - or doesn't fire if voltage stays below threshold. Whether neural computation is digital or analog is an important discussion in Neuroscience, and generally there is agreement over the fact that it is a mix of both - but that's a topic of discussion on its own.

Now that we have looked at generation of action potentials, how can we communicate it to adjacent neurons? The generated spike travels along the axon until it reaches the axon terminals - much like an electrical signal travels through a cable. Individual neurons are not directly connected with one another but separated by the extracellular space. In order to communicate a spike across this space, we need specialized structures right at the zone of contact between neurons: the synapses.

0.2.4 The synapse

In 1897 Charles Sherrington introduced the term synapse to describe the specialized structure at the zone of contact between neurons as the point in which one neuron communicates with another. The topic of synapses is a complex one: most of the computation of neurons actually happen at the level of the synapse. The synapse essentially is the element of the neuron that does most of the processing, and enables (or not) the action potential to happen. Synapses can be electrical or chemical, and allow excitatory or inhibitory dynamic into the post-synaptic neuron. The post synaptic neuron is simply the neuron that comes after the synapse - so if you have neuron 1 sending message to neuron 2, the neuron 2 is the post synaptic one, and 1 is the pre-synaptic one.

As shown in the figure above, the presynaptic neuron, i.e. the axon on the top, and the post-synaptic neuron, i.e. the dendrite, are separated by extracellular space, the synaptic cleft. The presynaptic neuron is filled with chemical neurotransmitters. Once an action potential arrives at the axon, these neurotransmitters are released into the synaptic cleft. The neurotransmitters fuse across the synaptic cleft and bind to the receptors of the postsynaptic neuron. This reaction causes a depolarization or a hyperpolarization of the postsynaptic dendrite. Whether the postsynaptic neuron is de- or hyperpolarized depends on the type of the synapse and some neurotransmitter dynamics. Synapses can either be excitatory, i.e. causing a depolarization, or inhibitory, i.e. causing a hyperpolarization, depending on the neurotransmitters they release upon activation.

A biologically-plausible model of synaptic transmission is given by the alpha function. It closely matches the shape of postsynaptic potentials that were measured during *in vitro* experiments. The alpha function is characterized by a gradual rise followed by a slow decay. For a single incoming spike, the response of the membrane potential is defined as $u(t) = te^{-\tau t}$. τ is the synapse's decay

¹⁶Digital, or boolean, logic is the fundamental concept underpinning all modern computer systems. Put simply, it's the system of rules that allow us to make extremely complicated decisions based on the aggregate of relatively simple 0s and 1s ("yes/no") elements. Neurons firing can be seen as 1s, and not firing as 0s.

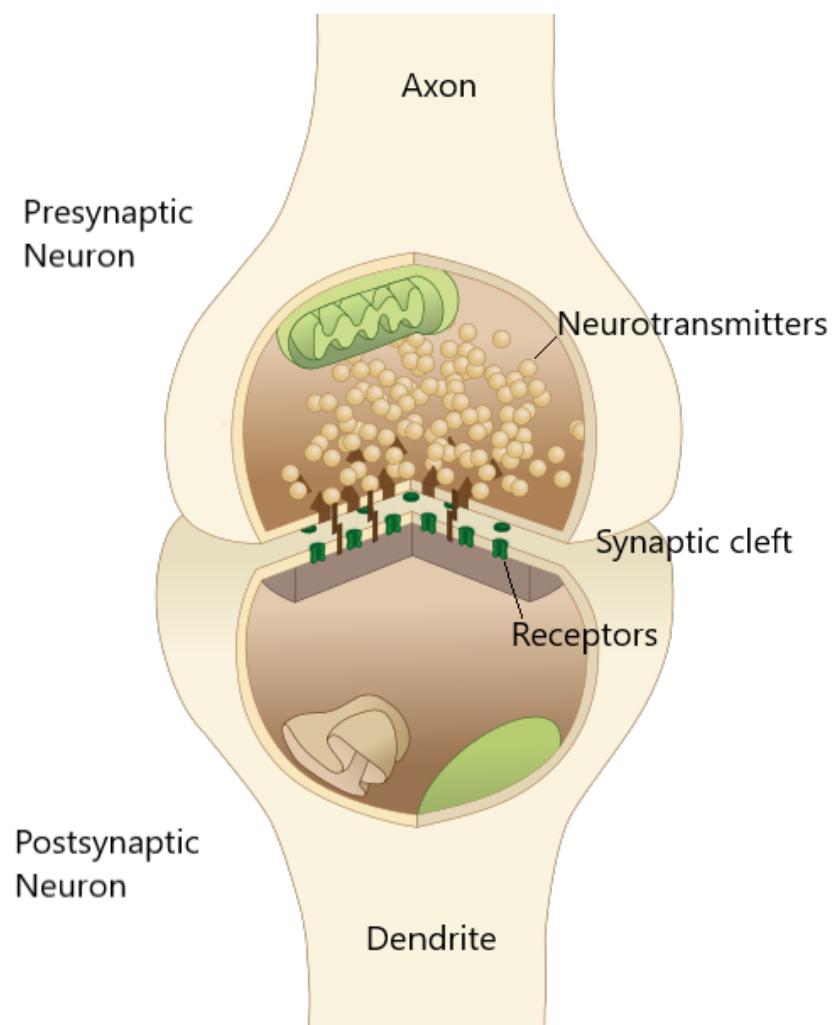


Figure 12: Schematic of a synapse. Adapted from Wikipedia.

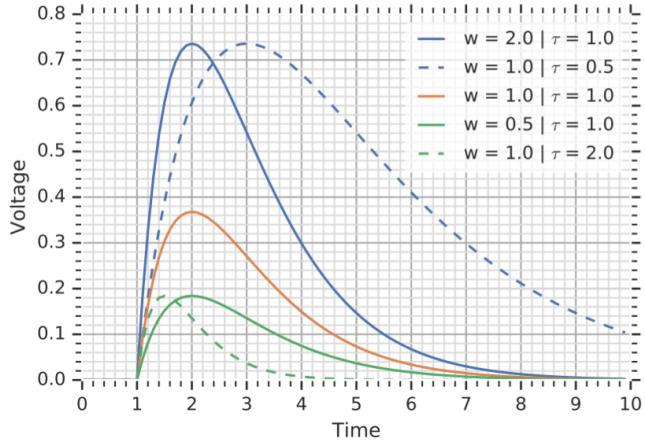


Figure 13: Alpha function as a biologically-plausible model of synaptic transmission (Taken from [Com+20]).

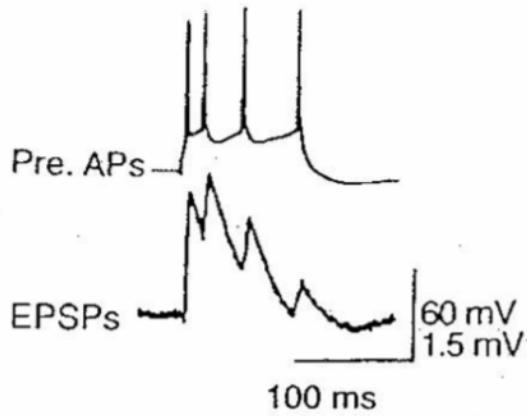


Figure 14: Excitatory post-synaptic potential (EPSP) in response to multiple pre-synaptic spikes. Adapted from Wikipedia.

rate which scales the output in amplitude and time. The amplitude is also shaped by the synaptic weight. Figure 13 visualizes the alpha function for different weights and decay rates.

What is important to remember is that chemical synaptic transmission is characterized by specific temporal dynamics directly correlate to the presynaptic neuron, as shown in figure 14.

Notice how the Excitatory post-synaptic potential (EPSP) changes in magnitude with the AP incoming from the pre synaptic neuron.

0.2.5 Network and Computation

One can easily see that these models of biological elements, though vastly oversimplified in my explanations, underlie very complex and intricate dynamics. In the brain, the complexity is several orders of magnitude higher, as all these elements interact with each other in a chaotic yet incredibly precise manner, which yields our ability to function. It is without surprise that many scientists chose to solely focus on understanding the emerging function from complex networks of the brain, and use the most basic of models of the brain's individual components. This simple idea has actually led to revolutionary progress, in various subfields. Mainly, Artificial Intelligence, with the Neural Network revolution. The concept is very simple: build a computational neural network, where neurons communicate with each other with a certain strength (synaptic weight for excitatory or inhibitory activation), and propagate the signal further along the network. These

extremely basic models of brain function are today revolutionizing the world. Other researchers have focused on trying to replicate models of neural networks on hardware, or in less obvious fashion than in traditional machine learning. This was the purpose of the *Human Brain Project* for instance, which fell short of so many of its ambitions. Carver Mead, the godfather of the field of Neuromorphic Engineering, envisioned copying some of the neural circuitry we understand in order to build faster and more efficient devices than the traditional computing counterparts. This has somewhat been successful, with the creation of the Silicon Retina by Misha Mahowald in the 1980s. Today, our understanding of the brain grows in complexity every day, and the idea of an elegant, simple and unified theory of neural intelligence seems more and more out of reach for humans. Nevertheless, simplified models have proven extremely effective for multiple real world applications, and it is worth looking at some of its basic concepts. It also is particularly relevant to Neuromorphic Engineering, as we will understand in due time.

0.3 Basics of Machine Learning

0.3.1 The Perceptron

As we have seen in section 0.2, the basic functionality of a neuron is to sum up the synaptic inputs it receives from its dendrites and generate an output if the inputs cross a threshold value. The first mathematical model of this functionality was introduced in 1943 by Warren McCulloch and Walter Pitts. The McCulloch-Pitts model is considered the first artificial neuron. Its structure is demonstrated in figure .

1 Neuromorphic Engineering: History, objectives and challenges

The field of Neuromorphic Engineering emerged in Caltech in the mid to late 1990's mainly through the work of Carver Mead et al. Carver is an eminent researcher who has enormously contributed to computing development and research. To understand what this field is all about, one should first look back at the birth of modern Neuroscience as well as the development of modern computing: both these fields witnessed major developments between the 1950s and 1970s which opened, as we all know, an ocean of opportunities for innovation in multiple fields.

1.1 Modern Neuroscience

The field of Neuroscience is certainly the most misunderstood of all "biological" fields. The main reason is that its serious study only started very late in scientific history, as there was no way for researchers to observe anything of interest inside a dead brain, and let's not even talk about the possibility of observing things in alive beings. So much that some early views regarded the brain as "cranian stuffing" - Egyptians believed that the heart was the seat of *intelligence*, so they literally took out some parts of the brain (through nostrils) after the death of individuals - yes, that was a long time ago, but that tells you something about how ignorant people were about the most complex human organ.

Fast forward to the 19th century, where it was accepted that the brain was human center for senses and intelligence, and where Neuroscience subsequently started to be recognized as an academic discipline in its own right.

1.1.1 Cajal and Golgi: the birth of Neurons

"Far from being able to accept the idea of the individuality and independance of each nerve elemnt, I have never had reason, up to now, to give up the concept which I have always stressed, that nerve cells, instead of working individually, act together [...]. However, opposed it may seem to the popular tendency to individualize the elements, I cannot abandon the idea of a unitary action of the nervous system." - Camilo Golgi, 1906

The m

- 1.1.2 Hodgkin and Huxley: The neural cell as a computational unit
- 1.1.3 Hubel and Wiesel Mountcastle: Experimental work on Neuron's Input to Output
- 1.1.4 Model of the brain as a complex system with detailed individual component behaviour

1.2 Modern Computing

- 1.2.1 The transistor
- 1.2.2 Perspective shifts and realizations: Von Neuman, Marvin Minsky, Feynmann and Carver Mead

1.3 Neuromorphic Engineering

1.3.1 Carver Mead and the Caltech Graduate Course

The train of events that led to this book began in 1967, when Max Delbrück introduced me to neurobiology. In 1982, Richard Feynman, John Hopfield, and I jointly taught a course on "The Physics of Computation." As a direct result of that course, I began to see how to undertake the creation of something resembling neural computation in the VLSI medium.

Carver Mead in Analog VLSI and Neural Systems Acknowledgment

"The train of events that led to this book began in 1967, when Max Delbrück introduced me to neurobiology. In 1982, Richard Feynman, John Hopfield, and I jointly taught a course on "The Physics of Computation." As a direct result of that course, I began to see how to undertake the creation of something resembling neural computation in the VLSI medium."

1.3.2 First Breakthroughs: Misha Mahowald and Silicon Retina

1.3.3 From research breakthroughs to deliverables

1.4 The rise of Machine Learning

1.5 Modern challenges and objectives

1.5.1 Energy considerations

1.5.2 Efficiency considerations

1.6 Challenges

1.7 List of complementary readings

- Von Neumann Silliman Lecture: The Computer and the Brain.¹⁷ This essay represents to me the birth of Computational Neuroscience. Von Neumann, saw and wrote (from his deathbed) the striking analogies that one could make between the brain and the computer, and introduced the world to the idea of studying them both at the same time.
- Analog VLSI and Neural Systems, Carver Mead¹⁸. The first textbook on Neuromorphic Engineering, by the creator of the field. This textbook is a marvel (see Chapter 0) and explains so many critical concepts of the field with stunning simplicity.
- Documentary on Misha Mahowald¹⁹. Great overview of the development of the field and the kind of work that emerged from the "Physics of Computation" course at Caltech.

¹⁷https://complexityexplorer.s3.amazonaws.com/supplemental_materials/5.6+Artificial+Life/The+Computer+and+The+Brain_text.pdf

¹⁸<https://www.amazon.com/Analog-VLSI-Neural-Systems-Carver/dp/0201059924>

¹⁹<https://www.youtube.com/watch?v=1wT1jUvwRLc>

- Giacomo's Porto conference ²⁰. Great introduction to the motivation, breakthroughs and challenges of modern Neuromorphic Computing by Giacomo - our teacher and a pioneer in the field.
- Shi-Chii's Conference ²¹. Similarly to Giacomo's conference, great introduction to the modern challenges and applications of the field.
- Event Based Vision: a survey. ²² Overview of one of the main applications of Neuromorphic Computing in a very comprehensive paper.
- Carver Mead IEEE guest paper ²³. This is an opinion paper from Carver Mead where he explains why he believes in the field having such a strong potential.
- Hubel and Wiesel paper ²⁴ and documentary video ²⁵. Their experiments are revolutionary and it is worth knowing what they're about.

²⁰<https://youtube.com/watch?v=cwQ8edHQFOA>

²¹<https://www.youtube.com/watch?v=tZM49YjUVDk>

²²<https://arxiv.org/abs/1904.08405>

²³<https://authors.library.caltech.edu/53090/1/00058356.pdf>

²⁴<https://authors.library.caltech.edu/53090/1/00058356.pdf>

²⁵<https://www.youtube.com/watch?v=8VdFf3egwfg>

2 The Essential Physics Behind the Transistor

As the main purpose of NE1 is to understand how to emulate brain function through elaborate transistor circuits, it is only necessary to grasp the intuition behind their intrinsic physical behaviour, and not the details of the physics, which would need a whole semester (at least) to cover properly. This is why we only focus on *grossly* explaining the concepts in a digestible manner, with only the details that will be absolutely necessary to understand later concepts. This chapter is subsequently devoted to presenting all the critical building blocks of the transistor, which will be covered in the following chapter.

Before starting this chapter, the student should be familiar with the following concepts:

- Basic atomic structure (high school level)
- All Electrical Engineering fundamentals introduced in Chapter 0.

2.1 Silicon: the magic semi-conductor

Silicon, which has given its name to the Silicon Valley, has joined since the late 50s the club of very important physical elements. Some greedy souls would even argue it's more important than Oxygen today: you can't make people pay for oxygen but you can make them pay for things that are built out of Silicon. Today, it's not just your typical sex-toy that uses Silicon(e)²⁶, but pretty much every electrical circuit. What makes Silicon special? How does it work? And how do we use it?

2.1.1 Prelude: Material Conductivity

Conduction is not difficult to understand. You can effortlessly walk through air. It's a bit more difficult to walk through water, but you can still do it if you have enough energy. Walking through a wall is merely impossible tho, except if you, like superman, are on Kryptonite. Same idea applies to materials, which are conductive, insulating or semi conductive. You need some energy to make a semi conductive material conductive, just like walking through water. Similarly, you can only manage to make a current flow through an insulator by applying an unusually high voltage. This is also why we say that insulators have infinite impedance (in practice, it just is extremely high rather than infinite) - following Ohm's law, you need an infinitely high voltage (think Voltage on Kryptonite) to have current flowing.

Moving on to one of the most important concepts to understand when it comes to this chapter (and transistors physics): Band Theory. This is the model that allows us today to understand what, at an atomic level, allows a certain material to conduct (or not) electricity, and the way it does so. This also allows us to understand how the conductive properties of the material evolve as a function of external factors, such as temperature or voltage for example.

Most of the following explanations are based on the excellent video "Band Theory (semiconductors) explained" from *PhysicsHigh* Youtube Channel²⁷.

First, we should remind ourselves that the atom is made of a nucleus with positive (and neutral) charges, and electrons around it in *discrete shells*. Note: if you're not already familiar with the structure of an atom, the concepts of shells etc., I suggest to look this up and come back here when you're familiar with it.

²⁶silicone or polysiloxane is a polymer made up of siloxane - silicones consist of an inorganic silicon-oxygen backbone chain (SiOSiOSiO) with two organic groups attached to each silicon center.

²⁷<https://www.youtube.com/watch?v=zdmEaXnB-5Q>

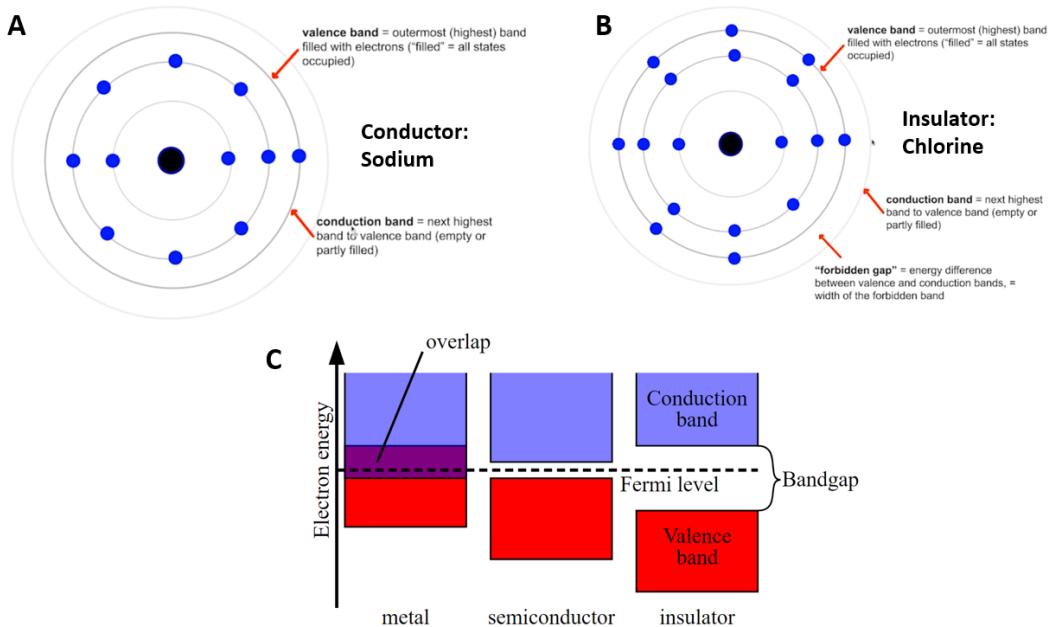


Figure 15: A - Structure of Sodium Atom - a conductor. B - Structure of Chlorine Atom - an insulator. C - Simplified and generic Band Energy Diagram of conductors, semi-conductors and insulators.

An atom such as Sodium has 11 protons and electrons, and are arranged as shown in Figure 15.A in shells - 2 in K shell, 8 in L shell and 1 in M shell. In this specific example, the M-shell is the outermost shell, which we can also call the **valence band**. On another note, the 11th elec is alone on the valence shell (where there are 7 other spots to fill), and it thus is *loosely held on that shell*, and is consequently rather free to move when a very small potential difference is applied²⁸. Sodium is thus a conductive material: it does not require much energy to have an electron moving around, which creates a current.

Now when we look at Chlorine in Figure 15.B, we have a different setting where the valence (outermost) shell is almost filled (7/8 electron space filled). This causes the electrons on that shell to be very tightly held together as opposed to what we saw before with Sodium. They are thus not free to move and thus cannot generate current. In order for them to move, we need to *provide enough energy* for some of them to "jump" to the next shell, where they eventually will be more free to move - the conduction band now corresponds to this next shell. To understand properly this, you need to go back to Bohr's theory that states that energy levels of electrons are quantized: it can have enough energy to reside in the M shell, and it can also have enough energy to reside in the next shell, but it cannot be in between where there exists a "forbidden gap". We therefore need to provide a very specific amount of energy for it to jump "the forbidden gap".

You guessed it, it's basically an in-between Sodium and Chlorine that makes up for a semi-conductor. We'll look specifically at Silicon in more details soon.

To sum things up: In a conductor, the valence band and the conduction band are overlapping: very little energy is needed to make electrons move around. In an insulator, the valence band and the conduction band are separated by a very significant energy gap: we need a huge energy supply (huge voltage for example) to make electrons in the valence band "jump" the "forbidden energy bandgap" to join the conduction band and finally become conductive. In a semi-conductor, the valence and conduction band are not initially overlapping, though we only need a relatively small amount of energy to have electrons make that jump, and most often, thermal energy (heat) is just enough.

²⁸Understanding why this is the case requires a lot more physics, so just accept this as being true for sake of simplicity.

Figure 15.C very simply represents the concept of what is called "Band Theory": a visual representation of energy levels of a given material. You can see that the valence band is overlapping with the conduction band in the case of a conductor, as we saw with Sodium. There is a small energy gap between the conduction band and valence band in the case of semi-conductors, and a very large energy gap in the case of insulators. Energy band diagrams therefore give you a visual representation of the amount of energy that is needed to make charges (holes and electrons as we'll see later) move from one energy state to another, and become conductive.

2.1.2 Thermal Consideration

Temperature is energy, and it affects properties of materials.

- **Insulators:** at 0 K the valence band is completely filled and the conductance band is completely empty. No charge transport can take place in either bands. In room temperature, there is still a negligible amount of charge in the conduction band.
- **Semiconductors:** at 0 K semiconductors are equivalent to insulators. The size of the band gap E_g is small in semiconductors (e.g. 1.1 eV for silicon (Si), 5 eV for diamond (C)). In semiconductors, the number of electrons available for conduction is significantly increased by thermal energy.
- **Metals (conductors):** the energy bands overlap, thus electrons and empty energy states are intermixed. At 0 K an applied electric field can generate current flow (electrons can move freely).
- The **Fermi Energy Level** is the difference between the highest and lowest energy of electrons at a temperature of 0 Kelvin.

Thermal Energy need also be looked at. We speak of **Thermal Voltage** U_t : it is the mean potential caused by the thermal motion of electrons. At *room temperature*, the thermal voltage is a relation between the charge q , Boltzmann's constant k and the temperature T in Kelvin:

$$U_t = \frac{kT}{q} = \frac{1.38 \cdot 10^{-23} J \cdot K^{-1} \cdot 300K}{1.6 \cdot 10^{-19} C} \approx 25mV = 1/40V \quad (11)$$

Note that this holds as 1 Volt = 1 Joule / 1 Coulomb.

2.1.3 Silicon: structure, doping and properties

I	II	III	IV	V	VI	VII	Zero
H							He
Li	Be	(B)	C	N	O	F	Ne
Na	Mg	Al	Si	(P)	S	Cl	Ar
K	Zn	Ga	Ge	(As)	Se	Br	Kr
Rb	Cd	In	Sn	Sb	Te	I	Xe

↑ | ↑
 Acceptors | Donors
 1 missing electron | 1 extra electron

Figure 16: Periodic Table. Silicon and Germanium are semi-conductors. Aluminum, Boron are Acceptor Impurity atoms: they yield free holes, Phosphorus and Arsenic are Donor Impurity Electrons: they yield free electrons. Adapted from Lecture Notes.

This section is extensively based on Jordan Edmunds excellent lecture series on semiconductor physics ²⁹. If you fail to understand what I write, I highly recommend checking out his explanation which go in greater depth into the topic.

²⁹<https://www.youtube.com/watch?v=OVnVN0vSXn0list=PLQms29D1RqeKGBEW8La2a7YuN54pSV4k>

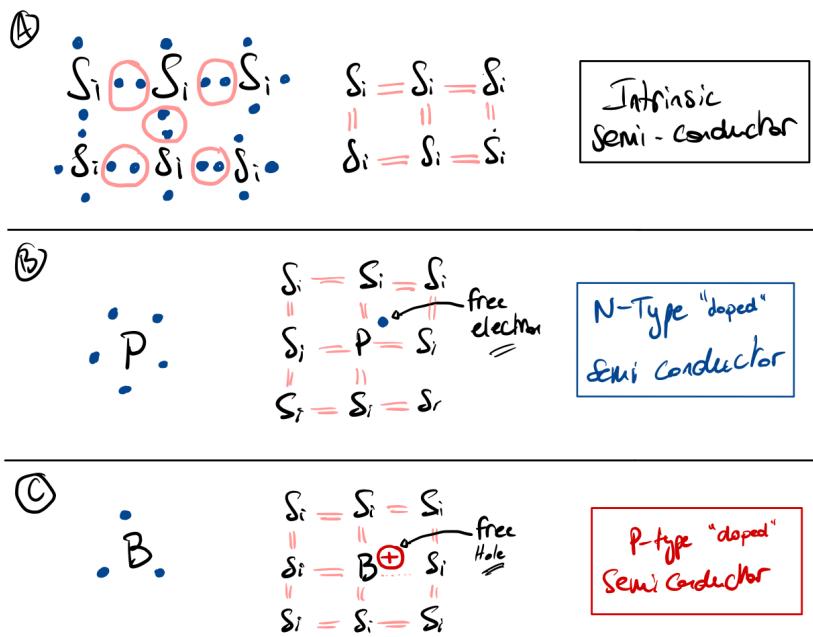


Figure 17: Intrinsic Silicon structure, and effects of doping to the structure. Blue dots are valence electrons, and red lines are "covalent bonds". Drawing by Yassine.

Silicon Structure Silicon is no exception to other semiconductors (and other materials) with atoms arranged in rectangular or diamond structure structures, known as *crystals* (see figure 17.A). These structures are defined and held together by the way the valence (outermost) electrons of the atoms are distributed. Silicon's atomic number is 14: it has 14 protons and 14 electrons, including 4 in its valence (outermost) shell. It takes two neighbouring valence electrons to form a *covalent bond*; this yields, in a silicon crystal, to each of the silicon atoms being bonded to 4 other silicon atoms (as shown in Figure 2.B). This is where the "diamond" analogy comes from.

Silicon Doping We've just presented the structure of *intrinsic* silicon (basically pure silicon). Intrinsic silicon happens to have a high impedance at room temperature (so it conducts, but not very well conducting), this cool video³⁰ reports an impedance for a 2cm piece of Silicon of 130.000 Ohms - that's quite a high impedance. We can significantly increase the conductivity of a semiconductor by **doping it with impurities**. Doping consists in a small fraction of the semiconductor atoms in the crystal structure to be replaced by atoms of a different element, which are obviously not chosen randomly. The atoms are of course electrically neutral, and they can be of two sorts: donors and acceptors. **Donor impurity** is an atom with a valence electron **more** than the semiconductor atom, and an **acceptor impurity** is an atom with a valence electron **less** than the semiconductor atom (see Figure 16. Let's look at the concrete examples of Phosphorus, and Boron.

As shown in figure 17.B, you can add a Phosphorus atom to silicon crystal. A Phosphorus atom has an atomic number of 15 and has 5 electrons in its valence shell (which are the blue ones represented in the picture). It will form traditional double covalent bonds with silicon with 4 out of its 5 outermost electrons. Will now be left that extra electron which does not have anything to bond with, and is just lying there, freely moving and looking for some mate to bond with. We consider that this electron has entered the conduction band. An important thing to realize is that phosphorus has just lost one of its electrons, and it still has 15 protons. So this phosphorus just turned into a phosphorus ion, and a positive one: P^+ . The issue with that charge is that it cannot be filled by just bringing in a new electron, because that electron would not

³⁰<https://www.youtube.com/watch?v=k12GMjtN8aAab> channel = MITOpenCourseWare

have any other electron to form a covalent bond with, which is why it lost an electron in the first place. To phrase it another way (and you'll understand in the next section why I need to phrase it another way): the positive charge accumulated in the phosphorus atom is **not** an available state which an electron can fill. As such, the positive charge in the phosphorus atom, the (*donor impurity*) is there to stay; however the negatively charged electrons are the ones moving around freely, thus justifying its name: the **N-Type Doped Silicon**. Before we move on to the next bit, note that we managed to add free moving electrons to crystal by doping it with phosphorus, but we actually do not change its overall charge! It still is a neutral piece of crystal, because it has the same number of free electrons than of positively charged phosphorus ions. Pretty cool huh?

Now let's look at what happens when injecting an acceptor impurity atom: Boron (see figure 17.C). Boron has atomic number 5, which means that he has 3 electrons in its outermost valence shell. When injected to Silicon, it will form traditional double covalent bonds with silicon on all 3 of its valence electrons. But here is the thing, Boron still has space for more electrons on its outermost shell, it still has one available state that could potentially be filled by electrons, we say that it now carries a **hole**. A hole is not a formal charge, but it behaves just like one: it is the absence of an electron in a particular place in an atom. So what you get is electrons hopping in this hole (and thus filling it). But when that happens, the electrons also leaves a hole behind it from the place it left. What you get is virtually a range of positively charged holes moving around freely in the **P-Type Dope Silicon** crystal.

Now that we understood the basic idea behind doping, let's look at some band energy diagrams to see how doping affects things.

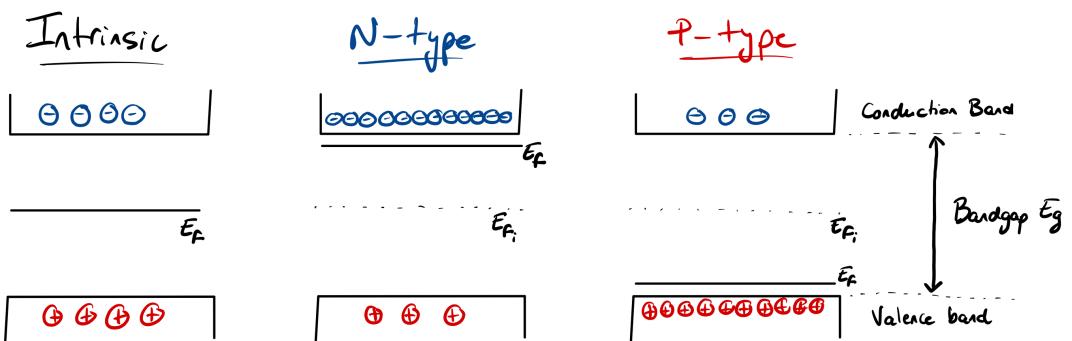


Figure 18: Band diagram of intrinsic, N-Type and P-Type silicon. E_{F_i} , in dotted lines, corresponds to the intrinsic silicon Fermi Level, which is the actual Fermi Level in intrinsic graph. For P-Type and N-Type, the effective Fermi Levels are at different positions than in the intrinsic case, shown in corresponding straight lines. Adapted from a bunch of graphs I saw which didn't really make sense to me, so I drew my own. Hope it's correct lol.

To me, this is the only necessary band diagrams to understand when it comes to intrinsic vs doped semi conductors. It is necessary to understand it in order to approach properly the PN-Junction, which we will look at in the next section. In general, you should remember that the Fermi Level is determined by the doping density of charge carriers (electrons and holes). In an intrinsic semi conductor, there are (schematically) as many holes as there are free electrons, and the Fermi Level is *not* in the conduction band, nor near it. Remember that the Fermi Level is the difference between the highest and lowest energy of electrons at a temperature of 0 Kelvin. That means, as we know, that an energy must be provided in order to make the material conduct electricity. When doping semi conductor in N-Type manner, we reach a level with a significantly higher level of free electrons, a lot of them already in the conduction band, thereby making the Fermi Level a lot closer to the conduction band, and thus more conductive for less energy provided. In the case of P-type, this is the opposite, and because the density of holes is a lot higher than the density of free electrons, the Fermi level is a lot closer to the valence band. However, one must remember that it is specifically *holes in the valence band* that can move freely and generate

a current flow. This is why at equivalent concentrations, N-Type and P-Type semi conductors have the same conductivity level ³¹.

Note: We can dope semiconductor to different levels. A weak P-type doping is denoted p^- and a very strong N-Type doping by n^{++} (you get the logic). If the semiconductor is so strongly doped that the Fermi level is within the conduction or valence band or very near the edge of one of these bands, such that a large fraction of the states at the band edge are occupied, its properties become similar to those of a metal and we speak of a *degenerate semiconductor*.

How is doping done in practice? You may wonder how does doping of silicon work in practice? The two main methods are:

- Thermal diffusion: Heat silicon in closed quartz tube, pump in some vaporized phosphorus or boron, and the dopants literally diffuse into the solid silicon
- Ion implantation: With an electron gun type of device, release small amounts of vaporized phosphorus, arsenic or boron, accelerate the resulting ions to a high velocity and ram them into the silicon.

At the dawn of semiconductors (from 1948 to 1960), diffusion was the primary method.

2.2 Understanding the PN junction and the Diode

Most explanations and graphics for this section are taken or inspired from the two brilliant explanation from *The Organic Chemistry Tutor* Youtube Video "What is a Diode", and *The Engineering Mindset* Video "Diodes Explained". The precise dynamics of what I am explaining could be studied over a whole semester, but understanding the basic idea is enough for what's needed from us in the exam. If you want more, refer to the Textbook, which will point you to appropriate electrostatics and semiconductor textbooks.

So, let's start. You might have heard of diodes before NE1 - they are semi-conductive devices that allow current to flow in one direction only. Ever heard of LEDs? Yes, they're diodes: Light Emitting Diodes.

³¹Take this with a grain of salt, I am really not sure whether that is the case in practice

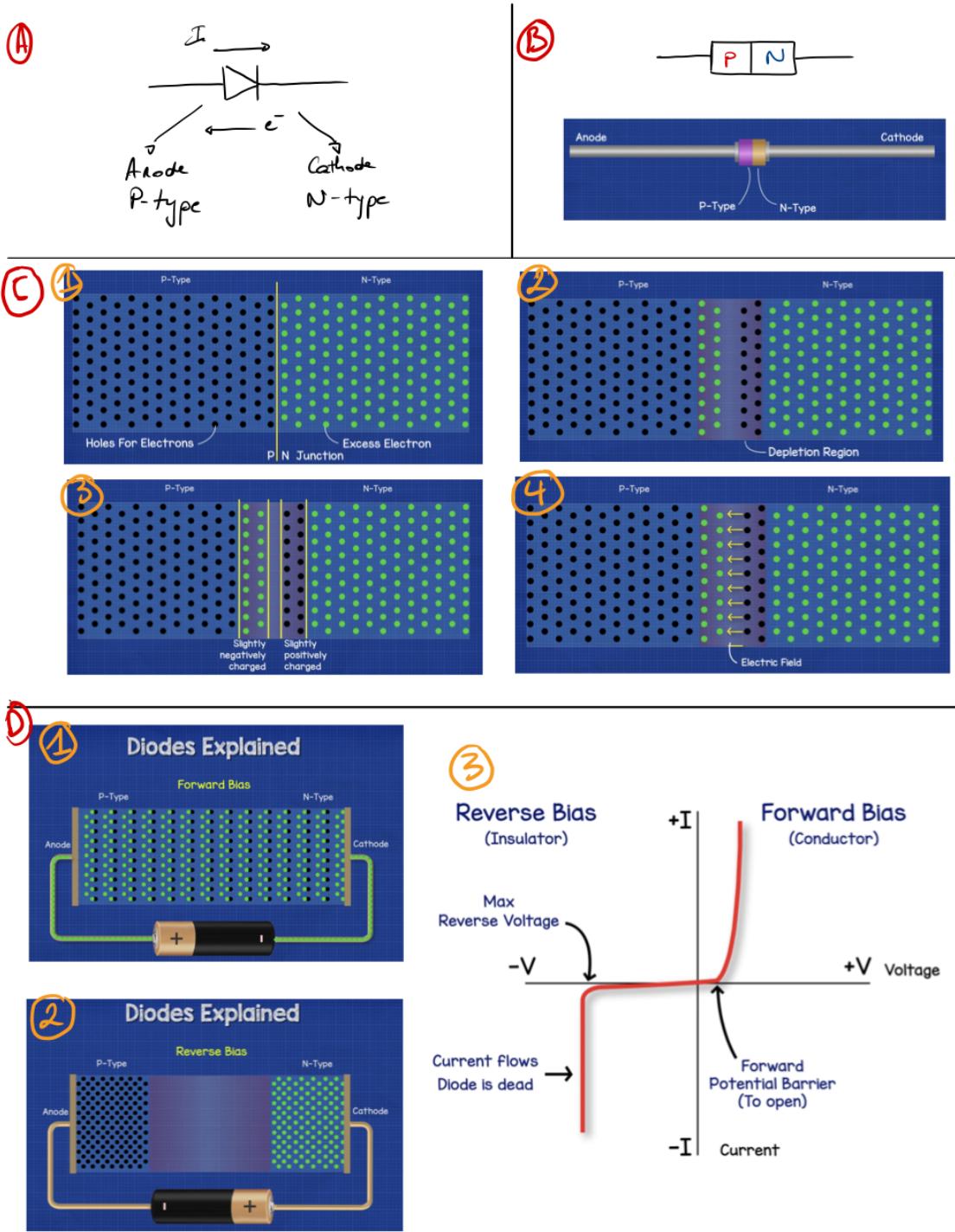


Figure 19: Understanding the diode. A) Schematic of diode in circuit with its anode and cathode. B) Schematic representation of the structure of a diode, and its *PN Junction*. C) What happens when sticking P-Type and N-Type semi conductor materials together. D) Forward and reverse current functioning principles in diodes. Adapted from multiple places.

Diodes are drawn in circuits as shown on figure 19.A. Current flows in the direction of the arrow, which means that electrons flow in the opposite direction (thanks again Benjamin Franklin). The "wall" at the tip of the arrow is there to remind you that current really only flows in one direction. So how does it do it? How does it allow current to flow only in one direction?

Diodes are essentially **PN Junctions**. A PN junction is what happens when you put P-Type doped semi conductor next to N-Type doped semi conductor (typically silicon). As shown in figure 19.B. the P-Type semi-conductor is known as the anode, and the N-type to an cathode. What happens when you stick P-Type and N-Type semi-conductors together?

This is the purpose of figure 19.C. As you already know, and as you can see in figure 19.C.1, the P-Type is filled with free holes and the N-Type is filled with free electrons. The junction between the two is what we call the *PN Junction*. 19.C.1 only shows you what you theoretically expect, but of course, as soon as you actually stick the two bits together, things start happening, which are shown in 19.C.2-4. The first thing to consider is that some of the excess electrons from the N-Type move over to the P-type side to occupy some holes - this occurs through a *diffusion process*. This yields the **depletion region** that you see in 19.C.2. This diffusive movement of electrons from N to P region causes some of the initially P-Type region to become slightly negatively charged and the N-type region to become slightly positively charged 19.C.3. An *electric field* is consequently created, in the opposite direction, which prevents further electrons from moving across (19.C.4). The depletion region is defined as *a region in semiconductor devices, usually at the juncture of P-type and N-type materials, in which there is neither an excess of electrons nor of holes*. This works with our figure and explanations!

In typical diodes, made of silicon, you should know that the potential difference between the slightly positive and slightly negative region is ≈ 0.7 Volts. This will become relevant soon.

Now what happens when you connect a voltage source to the diode?

This is the purpose of figure 19.D. If you connect the anode (P-type) to the positive end of your voltage source, and the cathode (N-type) to the negative end, you'll get (if the potential difference on your battery is $> 0.7V$) a nice current flowing in **forward bias mode** (19.D.1). Alternatively, if you reverse the power supply, for reasons that become very clear on 19.D.2, the holes are pulled towards the negative and electrons towards the positive, which causes the barrier to expand and (almost) no current to flow: we're in **reverse bias mode**. Yes, the barrier is similar to the band gaps we talked about before - and this is why in reverse bias mode, the diode acts as an insulator. We now know how diodes only allow current to flow in one direction. Pretty cool right? The stronger the reverse bias, the larger the surface area of the depletion region! This will be particularly relevant in the photodiode chapter. Now to enter a bit more details, which are shown in 19.D.3: current really starts flowing when you apply a voltage (in the right mode) superior to some "forward potential barrier", which in the case of silicon is $\approx 0.7V$. From then, current flows. Now when we apply a voltage in reverse bias mode, you can see on 19.D.3 that the current flow is not exactly 0 (because it really never is anyways). If you apply a strong enough voltage in reverse bias (remember, voltage on Kryptonite), you eventually will force a current, and probably damage your diode by the same occasion (just like Superman would break the wall and not simply go through it). In general, you always have both a forward and reverse current component, that's just how electron dynamics work, though depending on the voltage you apply, one is several orders of magnitude more important than the other, which is why we neglect it.

What does this give in Band Energy Diagrams? You thought you could escape it didn't you? Don't worry, we'll make it simple.

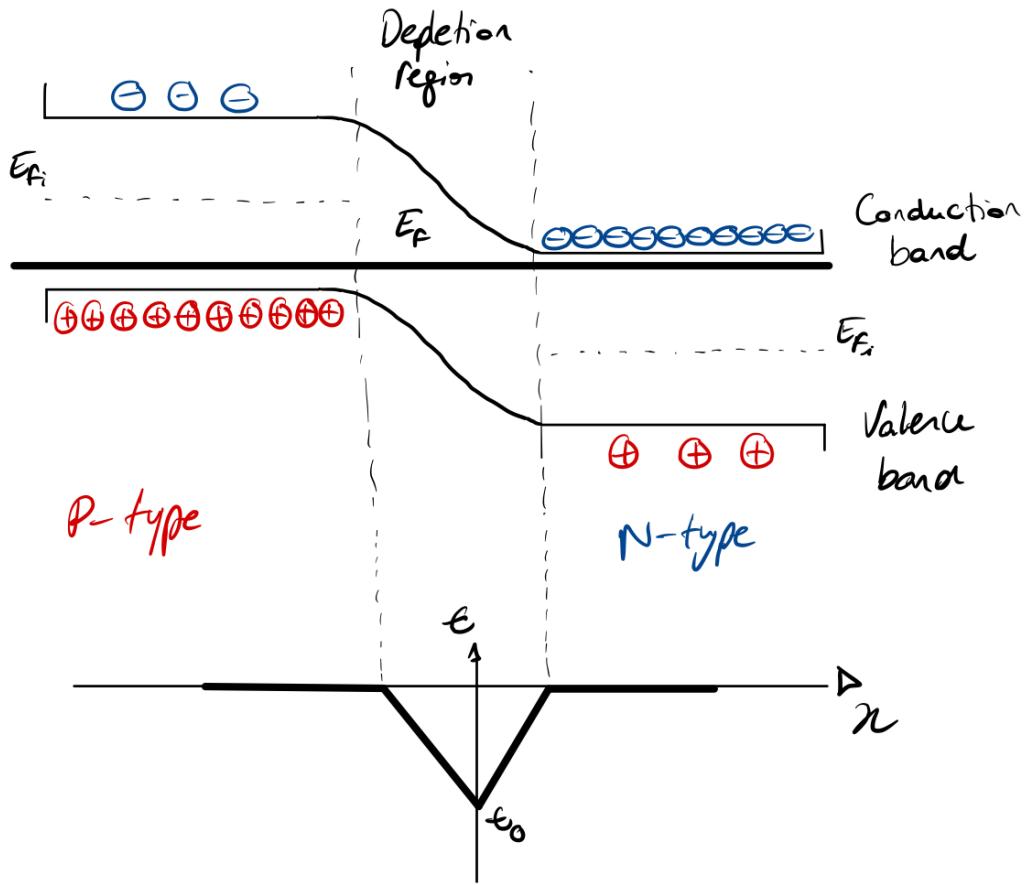


Figure 20: Band diagram of the diode, and its corresponding electric field ϵ in the depletion region. Electric field is negative as it, by convention, is from positive to negative, and the arrow points from negative to positive. Yassine drawing, attempting to simplify the Textbook drawing.

2.3 MIS Capacitance Structure

We are now reaching the last part of what we need to understand in physics before getting to the transistor. The *Metal-Insulator-Semiconductor* (MIS) structure consists of assembling some conductor and some semiconductor, separated by a thin insulating layer. The most common version of the MIS structure is the MetalOxide-Silicon (MOS) structure, where the ‘oxide’ is in most cases silicon dioxide (SiO_2). The MOS structure and the PN junction diode are the building blocks of today’s most widely used type of transistor, commonly known as Metal-Oxide-Semiconductor Field-Effect Transistor (MOSFET). This is the transistor we will study in the next chapter.

MOS capacitor structure

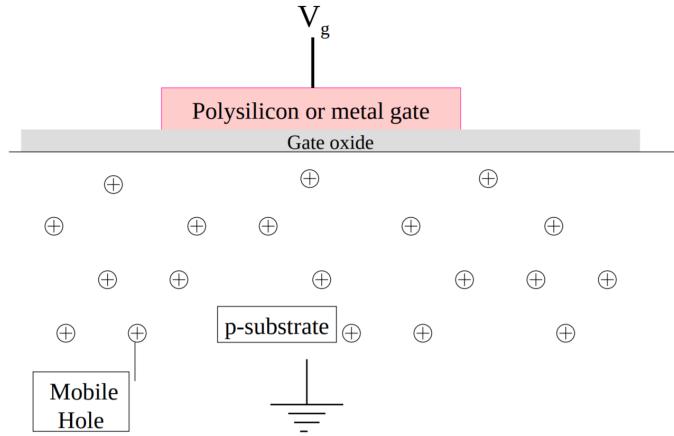


Figure 21: MOS Capacitor structure. A thin layer of insulating *Gate Oxide* separates two conductive material: a polysilicon metal gate and a doped semi conductor *p substrate*. Adapted From Lecture Notes.

As shown in figure 21, a thin layer of insulating *Gate Oxide* separates two conductive material: a polysilicon metal gate and a doped semi conductor *p substrate*. The p-substrate is simply p-doped silicon, which is at a potential of 0V in this case, as can be seen on the figure (it is connected to ground). Because of the capacitive structure, interesting things start to happen when we apply a non-zero voltage on the metal gate, which is what we need to look at. Typically, imagine that when applying a negative voltage on the metal gate, negative charges accumulate on the gate and positive charges accumulate on the other side of the insulating layer in the p-substrate. Conversely, when applying a positive voltage on the gate, positive charges accumulate on the gate and negative charges accumulate on the surface of the insulating layer of the p-substrate (it's more complicated than this as we'll see, but this should just help you build the intuition). Essentially, very distinct dynamics are obtained as a function of the applied gate voltage - this is critical to understand the transistor. Let's look at all the different case scenarios and their implications ³²:

³²We'll only take the example of p-substrate, but the inverse holds when working in n-substrate.

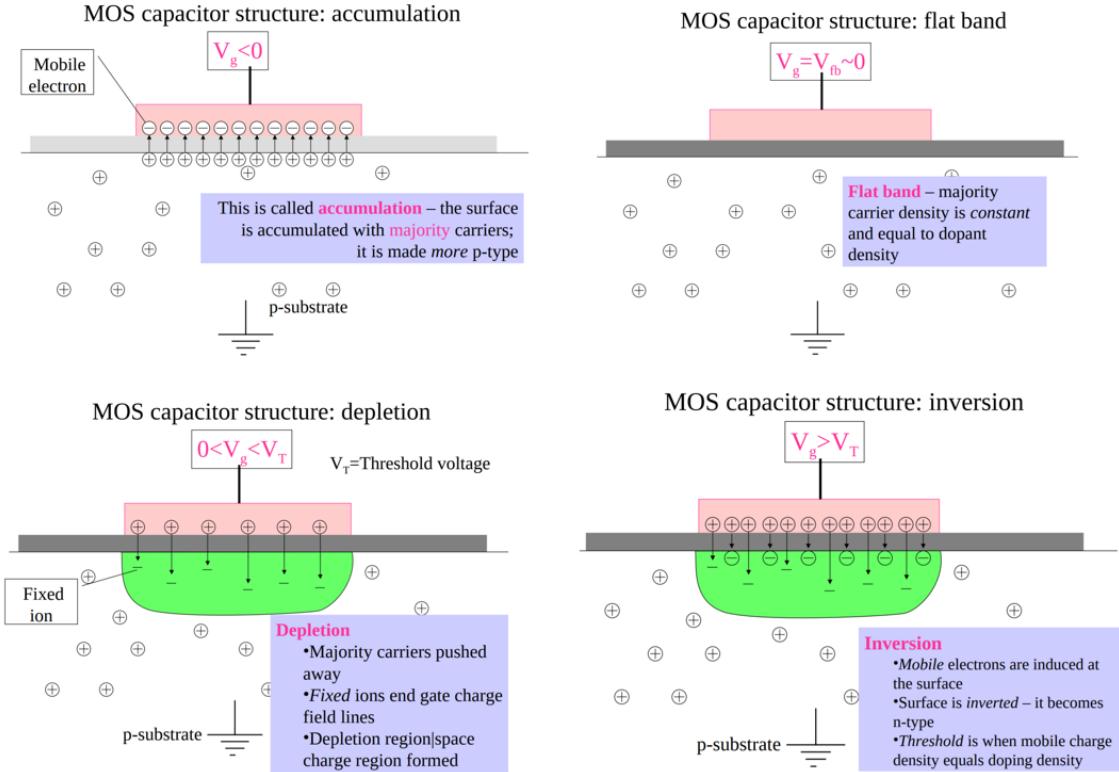


Figure 22: MOS Capacitance and its different operation modes. Adapted From Lecture Notes.

Accumulation Accumulation happens when a charge on the gate attracts a lot of **majority** carriers on the surface of the semiconductor underneath it. For a p-substrate channel (as shown in figure 22), a negative voltage on the gate means mobile electrons on the gate. These electrons attract positive charges on the semiconductor surface underneath it, leading to accumulation of them. Since positive charges (holes) are the majority carrier in p-substrate, we call this situation accumulation. It is made even more p-type.

Flat Band In an ideal MIS diode, with no bias applied ($V_g = 0$), the work function of the metal and the semiconductor are the same. The Fermi levels line up and the energy bands in the semiconductor are flat. In Flat-Band Condition, the majority carrier density is constant and equal to the dopant density.

Depletion For a p-substrate channel, if we apply a positive *subthreshold*³³ voltage on the gate (positive charges), we push away positive majority carriers on the semiconductor surface. A depletion region is created where only **fixed** negative charge carriers remain. This is the same depletion region we saw in the PN-Junction; but instead of being formed naturally, you create it by pushing the majority carriers of the p-substrate away.

Inversion Eventually, by increasing the positive voltage on the gate and pushing away majority carriers from the p-substrate away from the gate, we progressively start having **mobile** minority carriers (electrons here) accumulating at the semiconductor surface - this is called *inversion*. Inversion happens when the surface is inverted: *it becomes n-type silicon rather than p-type*. The precise gate potential at which this happens is known as the **threshold voltage** and is what distinguishes depletion from inversion mode. The threshold voltage is thus the exact voltage value at which mobile charge density equals doping density - when we go beyond, mobile charge takes over!

³³explanation on this coming in the Inversion paragraph - don't worry, it will make sense

These dynamics are summarized in the following figure:

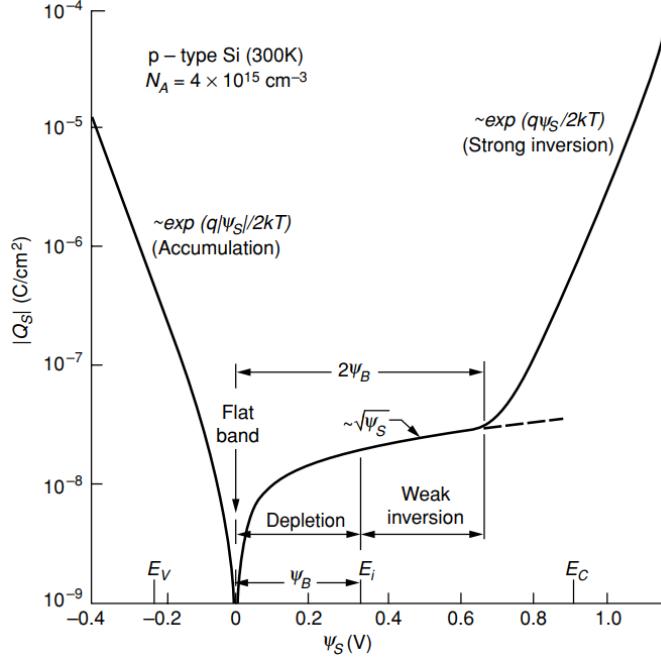


Figure 23: Dependence of the area charge Q_s on the surface potential for p-type silicon with acceptor density $N_A = 4 \cdot 10^{15} \text{ cm}^3$ at room temperature. Adapted From Textbook.

Here we can introduce quantification of charge. We define Q_g as the amount of charge accumulated at the gate. What one should pay attention to in this figure is how the surface potential changes as a function of Q_s . Q_s is simply $-Q_g$, which is, like Q_g , directly related to the gate voltage V_g . Of course, this dependence varies with different doping concentration and other such factors, but the nature of the relationship remains the same. Now we can even draw some equation to summarize this point:

$$Q_g = -Q_s = Q_d + Q_i \quad (12)$$

With Q_d charge on the depletion layer and Q_i charge on the inversion layer. Essentially, this yield a sort of capacitive model - the MIS Capacitance Structure - where we can consider charge at different points in the structure, and distinguish two capacitors: one actual capacitor with the gate and the inversion region, and another virtual one between the depletion region and the substrate. We can say that the one between the depletion region and substrate is virtual as there is no actual insulators between the two charged regions, but there clearly are opposite charges on both ends, just like a capacitor.

All the important elements subsequent from what we've just described are presented in the following 4 slides, which are shown on figure 24.

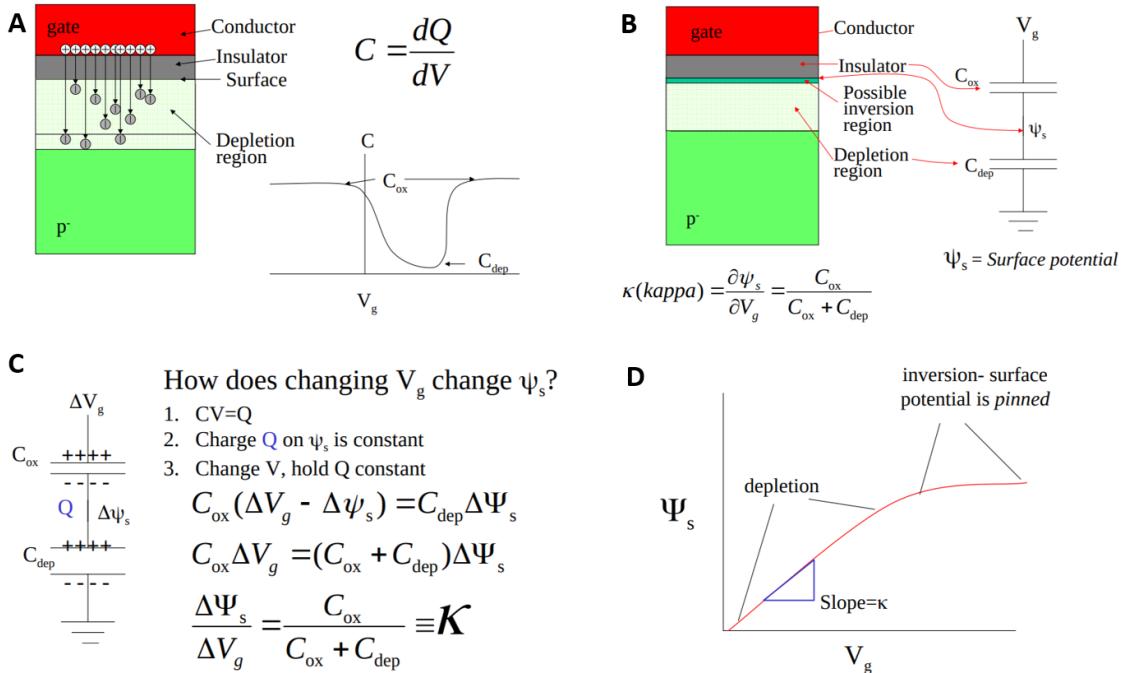


Figure 24: A) The depletion capacitor. B) Influence of the gate on the surface potential. C) How does changing V_g change ψ_s . D) Change in κ below and above threshold. Adapted From Lecture Notes.

Let's unpack all of this.

A) The depletion capacitor Let's remind ourselves that $C = \frac{dQ}{dV}$. This means that the more charge is accumulated per change in voltage, the higher the capacitance. We've described in the depletion regime that mobile charges are **pushed away**, only to leave **fixed ions**. This means that capacitance is **reduced** in the depletion regime! Indeed, it doesn't bring in new charges as we increase voltage, but actually pushes charges away - so dQ is small for a given dV , at least a lot smaller than for accumulation or inversion regime. This justifies the graph we see on 24.A. Now when we go above threshold voltage, in inversion regime, remember that the voltage is so strong that mobile charge carriers are attracted to the surface. Change in voltage yields important change in charge carriers: capacitance is big again! We subsequently differentiate between Oxyde capacitance C_{ox} and depletion capacitance C_{dep} .

B) Influence of the gate on the surface potential On this figure, we see the full idea of the MIS capacitance structure, with the gate voltage, the oxyde capacitance, the surface potential, the depletion capacitance and the substrate - connected to ground. The surface potential changes as a function of the gate voltage - we define the change in surface potential ψ_s as a function of the change in gate voltage with a value of important significance: κ .

$$\frac{d\psi_s}{dV_g} = \frac{C_{ox}}{C_{ox} + C_{dep}} = \kappa \quad (13)$$

Because this rate of change is different depending on the operating regime (depletion or inversion), it will have a different value as a function of the region. We will see this clearly in 24.D. The math behind this will also become clearer in 24.C.

C) How does changing V_g change ψ_s Here, we are trying to understand the equation derived previously, that is, why $\frac{d\psi_s}{dV_g} = \frac{C_{ox}}{C_{ox} + C_{dep}}$. The basic assumption to understand this is that the charge doesn't change: whatever charge appears on C_{ox} need also be on C_{dep} , simply because of charge conservation. No charge can be created! From this, understanding the equation on the

figure is (relatively) straightforward. Now one intuition to remember: if C_{ox} is big, κ will be close to 1

D) Change in κ below and above threshold κ is the slope of ψ_s vs V_g ! It is constant below threshold and experiences a shift when approaching the threshold to become again constant, and close to 0, above threshold. This is because, in inversion, the surface potential is pinned - so the change in surface potential is very small when changing gate voltage.

Concluding thoughts on MIS Capacitance: If this is not perfectly clear for you, you're not alone - this is one of the trickiest concept about device physics and really the whole module. It also happens to be one of the most important one to understand transistors dynamics. The issue is that abstractions and simplifications have limits: the dynamics described here are very complex and need substantial physics to understand clearly. But because this really is not the topic of Neuromorphic Engineering 1, it is taught rather quickly. In any case, there are some absolutely critical points that you should remember and be fairly okay with about this:

- What MIS Capacitance stands for
- How changing the voltage at the gate influences what happens in the semiconductor substrate on the other side of the insulating layer
- Qualitatively describe the difference between Accumulation, Flat Band, Depletion and Inversion Regimes.
- Understand the very important concept of the threshold voltage, and what specifically happens when we cross this threshold.
- The two different capacitors in the MIS Capacitance: Oxyde capacitor and Depletion Capacitor
- The idea behind κ , and how it relates surface potential to gate voltage (equations + slope).

2.4 Test yourself

You should know and understand the following things:

- Structure of Silicon.
- Difference between conductors, semi conductors and insulators on an atomic perspective, and with some intuition about the band energy diagrams.
- The idea of doping, how that changes conductivity and band energy diagrams.
- The idea behind the PN Junction and the dynamic it creates (depletion, electric field etc..).
- How the diode works, and the principle of forward vs reverse bias mode.
- The idea behind the MIS Capacitance structure and its operating modes as a function of gate voltage.
- What κ represents and how it relates the change in surface potential vs gate voltage.

3 Transistor Operation

3.1 Building an Intuition of the Transistor

We have in the previous chapter presented all the necessary information to understand the basic structure of NE1's rockstar: the transistor. Specifically, we'll be looking at the *Metal-Oxide-Silicon-Field-Effect-Transistor*, commonly called *MOSFET*, which is a particular type of Transistor, extremely common in industry. There exists other types of transistors that we don't look at in this course, such as the BJT (Bipolar Junction Transistor), as well as many others. The currents in these devices comprise either positively-charged holes, negatively-charged electrons, or both holes and electrons. The BJT is called a bipolar device because the current in the transistor consists of both types of carriers, electrons and holes. The MOSFET is called a unipolar device because the current has only one type of carrier, either holes or electrons. All transistors have at least 3 terminals (often 4). Though they all exhibit similar behaviours, some differences make them more useful in some applications. The following section is mostly about developing an intuition about the transistor structure and function. We will get into the specifics of operation in the next section and derive the precise equations that best describe its behaviour.

The main purpose of transistors is to control the flow of current from one node based on the potential at another node: by controlling voltage, you control current. Once again, let's start with some fluid analogy to build intuition on the transistor.

3.1.1 Understanding the transistor idea with Hydraulic Analogy

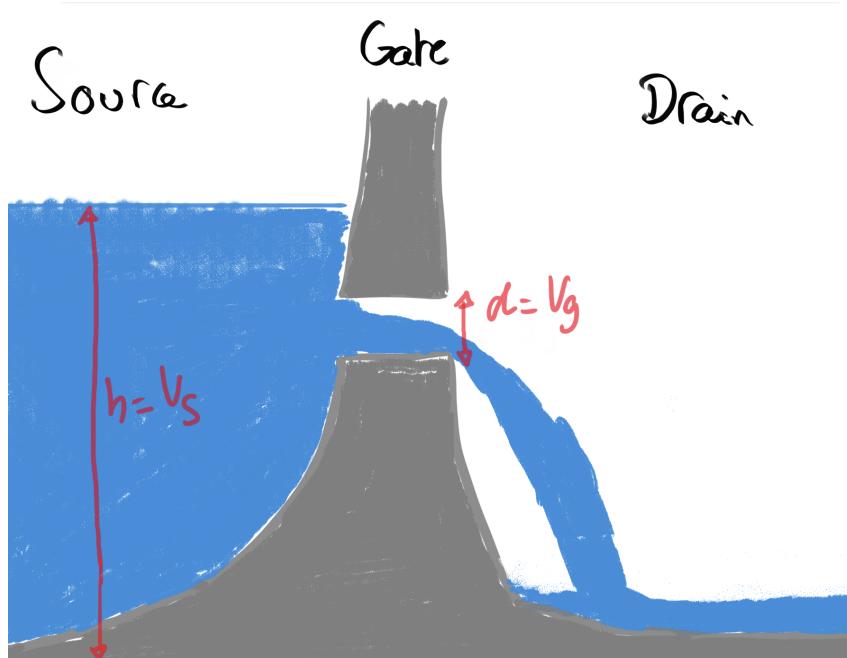


Figure 25: Water analogy to understand the transistor. Water in a reservoir (source) separated from another bit, called the drain by a gate. We can open the gate by *lifting* the upper part, to a certain height, which will influence the current flow. Yassine's drawing, inspired by a bunch of drawings I saw online.

I never managed to find the right drawing for what I was trying to convey, so I drew my own. Things are pretty self explanatory on the figure: there is water in some sort of reservoir, that we can call **source**, and no water in the **drain**. Let's imagine that the water in the source doesn't run out, i.e. each molecule that eventually flows out is brought back through some external supply not represented here - so h stays fixed. Now, remember from chapter 0 that because water is at a certain height h compared to the ground (where the drain is), there is some *potential difference*

between the source and the drain - this is a very important point. There is also separation between the source and the drain, called the **gate**. We can now bring some energy to open this gate (as it needs to be lifted - I know my drawing isn't great, but just imagine you need to lift the upper part to open the gate) and open it up to a certain level d . The level at which we open it will determine the extent by which water flows from the source to the drain. Now imagine that the drain gets full up to the same level as the source - even if the gate is opened, there is no net flow between the source and the drain. So we need **both** an opening of the gate *AND* a potential difference between the source and the drain in order to generate flow.

The way we open the gate also has an impact on the way water flows. Imagine that d is small compared to h , then you have a small flow. The more you open the gate (increasing d), the more things flow. The water flow (in our analogy) is **exponentially** dependant on the gate opening. But if we open the gate to some high $d \geq h$, well water flow does not really depend on d anymore³⁴! Past this *threshold*, the gate is opened and water flows independant of how much bigger our opening is.

It should be clear the kind of dynamics that we're trying to put in evidence here: when we have a potential difference between the source and the drain, water wants to flow. This water will flow if we open the gate, and the nature of the flow will change depending on how large the gate opening is. This is the basic idea of a transistor.

One thing that needs to be added to the analogy: water evaporates. It is possible when the gate is open that water from the drain gets into the source, but this is in such small quantity that it is often negligible - though it is still there! The higher the potential difference between the source and the drain, the lower the chances that water evaporates from the drain to the source. This is the idea of forward and reverse flow, which is important in transistors as we will see. When the potential difference between source and drain is high, forward flow is overwhelmingly dominant, when the potential difference is very small, the reverse flow should also be considered to some extent, in particular conditions.

3.2 MOSFET Structure and basic function

The purpose of this section is only to apply the hydraulic intuition to the transistor, by studying its structure and basic function. The details of the operation of transistors are significantly more complex than what we will cover here, and most of the necessary functional details will be covered in the next section.

³⁴Remember that there is a constant supply of water in the source so that h stays constant

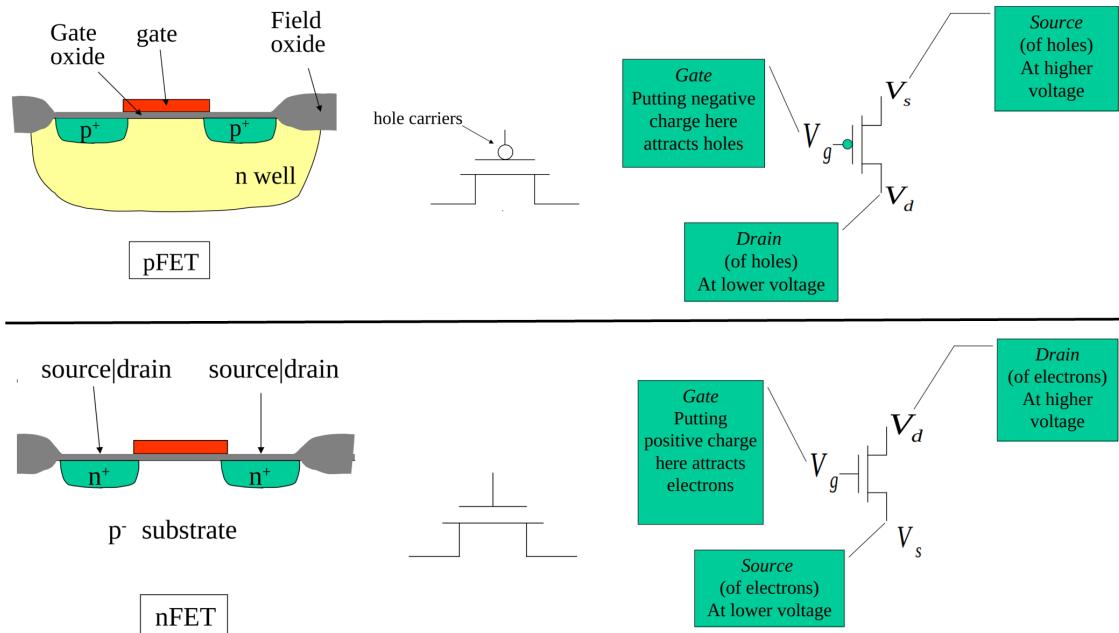


Figure 26: Essential structural and schematic description of the two transistor type: the pFET and the nFET. On the left, structure. In the middle, circuit schematic. On the right, principle of the 3 operating points. Adapted from lecture notes

There are two distinct types of MOSFETs: the pFET and the nFET. While they use the same principle of operation, some basic differences should be noted. There exists also other type of practical operation differences that we will look at later.

Both the pFET and the nFET have 4 terminals. That is, they both possess a **source**, **drain**, **gate and bulk**. The source and drain are pieces of doped silicon - p-type doped in the case of pFETs and n-type doped in the case of nFETs. You can see in the schematic p^+ and n^+ for the source and drain terminals, with the + indicating the doping concentration. Typically, both source and drain have the same doping concentration, but more on this later. In both the pFET and nFET, the source and drain are in contact with a **bulk, well or substrate** (take them as representing the same thing for now) of opposite type. That is, the n-type doped silicon source and drain in an N-FET are in sitting on p-type doped silicon well. This should ring a bell: we're dealing with PN-Junctions!

We do **not** want to have current flowing from the source and/or drain to the well, we actually want it to stay still (for now) - so we **reverse bias** the PN-Junction! That is, in a nFET, we apply a *lower* potential on the p-well than on the source and drain, and vice versa in a pFET³⁵. You should see clearly why this yields a reverse bias function, and a subsequent **depletion region**.³⁶

Now what about the gate? The gate is a piece of conducting metal, and it is separated from the source, drain and well with a **gate oxide**. This is an insulator material. This should also ring a bell: conductive materials separated by an insulator - we're dealing with some capacitor - a MIS capacitance structure!!

On the nFET diagram, you may have noticed "source/drain" on both terminals. This also holds for the pFET. Essentially, the transistor is a perfectly symmetric device, and both terminals can be used as source or drain. Now let's look at what exactly is the difference between source and drain.

³⁵No potential difference between source and well is actually okay, so you either need lower (or higher in the case of pFET) potential difference or no potential difference at all to keep this reverse bias.

³⁶If you don't see this clearly, I suggest reviewing the section on PN-Junctions, as well will continue building from this in the next bits!

Source and drain: what's the difference? Remember from the hydraulic analogy that our source and drain of water had some potential difference. This is exactly what happens in the transistor, and this potential difference will determine what is the source and what is the drain. Things have the same logic, but in reverse, for the pFET and nFET: this is showed on the right hand side of figure 26. In an pFET, the source is at the node at the highest voltage. We call this V_s for *source voltage*. The *drain voltage* V_d is at a lower potential. Remember that this is not the same as the doping concentration. If we choose to apply the highest voltage on the right terminal of the pFET, this becomes the source and the other becomes the drain, and vice versa. Again, it is symmetric and its operation depends on the voltage difference between the source and drain. This voltage difference has a specific name: **drain to source voltage**, and is note V_{ds} it is simply the difference between the two: $V_d - V_s$. In the case of the nFET, it's exactly the opposite. The source is at the lowest potential and the drain at higher potential. This should make sense intuitively: in a pFET, charge carriers are positive holes, there are more present in higher voltage points, which thus make them the source - they travel to the drain at lower potential. In a nFET, charge carriers are negative electrons, there are more electrons in the lower voltage points, which thus make them the source - they travel to the drain at higher potential.

Circuit diagram Both diagrams of the pFET and nFET are built around the same structure, except that the pFET has a small circle between its gate and the line. This symbolizes the holes, as opposed to electrons, which are the charge carriers in a pFET (as the source and drain are p-type doped silicon).

The gate Now that we conceptually explained the difference between the source and the drain, we need to look at the gate. Remember from the hydraulic analogy that the gate is what controls the flow of water, provided we have a potential difference between the source and the drain. The same principle applies in transistors. The dynamics are best understood schematically - and keep in mind that this is just to build an intuition!

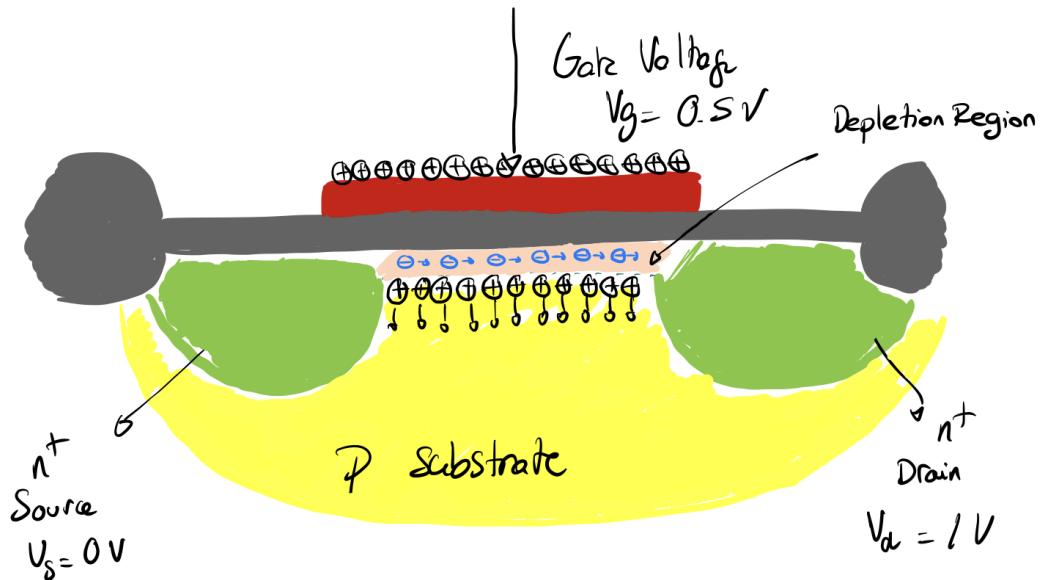


Figure 27: Basic principle behind current flow when applying gate voltage. This drawing is **NOT** completely physically accurate - it only serves the purpose of giving you an intuition of how current flows in a transistor. Yassine's drawing, adapted from a few figures in the lecture notes.

Disclaimer: This drawing is **NOT** completley physically accurate - it only serves the purpose of giving you an intuition of how current flows in a transistor. Anyways, let's look at it to understand how transistors work. I drew a N-FET with a V_{ds} of 1V - so potential difference between the source and drain is present. We now apply a gate voltage $V_g = 0.5V$. Because of the positive voltage, we

have a positive charges that accumulate on the gate. This naturally repels the positive charges on the p substrate (because alike charges repel each other) and creates some kind of corridor virtually of positive charges, where negative charges start to accumulate. This corridor suddenly allows a flow of charges through diffusion between the source and the drain as we have a significant drain to source potential difference, and thus a current. This is because, in this region, we are not anymore in a standard equilibrium PN Junction configuration anymore. This region, drawn in orange on the figure is the **depletion region**. This is *grossly* what happens when current flows. Now in reality, we must go back to our previous explanations on MIS capacitance, and we'll see that the nature of the flow is different if we are in the **depletion** regime or the **inversion** regime. In the depletion regime, current is driven by diffusion, whereas it is drive by electric field in the inversion regime. More on this in the following sections.

3.3 Understanding Sub-threshold Current using Boltzmann Distribution (Written by Wencan Huang)

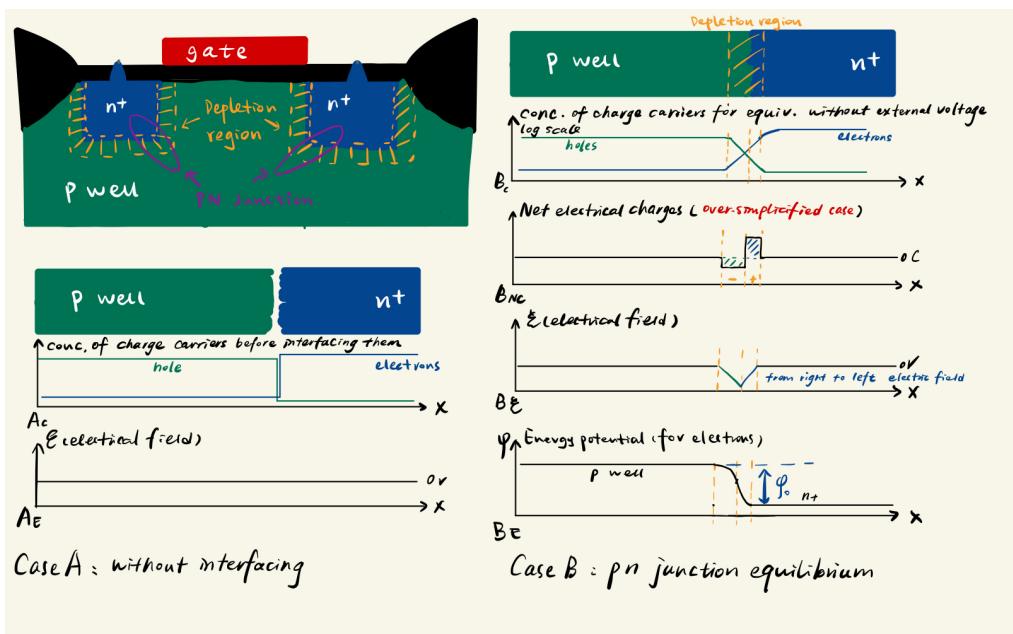


Figure 28: An illustration of a nFET transistor without attaching external voltage to its terminals. And case A and case B help explains how depletion region between highly doped n type node with p well is formed, and what properties it has, e.g., electrical field and energy potential. The depletion region is then oversimplified to be considered as have exactly no free charge carriers for further calculation. Figure drawn by Wencan Huang

nFET transistor without connecting to voltage supply Without loss of generality, detailed physics and math expressions are deliberately omitted and any curious readers can refer to semiconductor physics textbook or the Caltech online courses. Knowing nothing about those details is totally okay for succeeding in this course, however, it is necessary to understand the general idea presented in this section.

When there are no external voltage difference attached to nFET transistor's terminals, the depletion region between p well substrate and two n type nodes is formed due to diffusion and reaches equilibrium due to the electrical field formed during diffusion as is indicated in the figure 28 for electrical field in plot B_e .

For simplicity, take n type side for consideration and the same analysis can be followed for the p well side. The diffusion effect on the n type tries to diffuse n type free charge carriers - electrons - among the whole p-n junction. However, as electrons diffuse into p well side, they will leave net positive charge in n type side through fixed positive ions which lose valence electrons, and

cause net negative charge in p well side through contributing electrons to group III element. As a result, an electrical field is formed counterbalancing the diffusion of electrons towards p well side, and the equilibrium of diffusion is reached as drawn in figure 28 for concentration of free charge carriers in plot B_C .

Due to the effect as mentioned in the paragraph above, with oversimplified mathematical models, we assume that within the depletion region, the charge density is constant as shown in plot B_{NC} . Based on this assumption, we gain the consequent plots for electrical field and energy potential for electrons. In general, for silicon transistors, the potential gap ϕ_0 is around 0.7V.

nFET transistor connecting to voltage supply without considering V_g yet

For neuromorphic engineering I, what we care most is transistor's operation in sub-threshold region, i.e., $0 < V_{gs} < V_{thr}$ add the hyperlink to related sections here and ensure the symbols are consistent. Besides, a nFET transistor is only used with source-p well and drain-p well reverse biased. And the rest part will only consider under such circumstances.

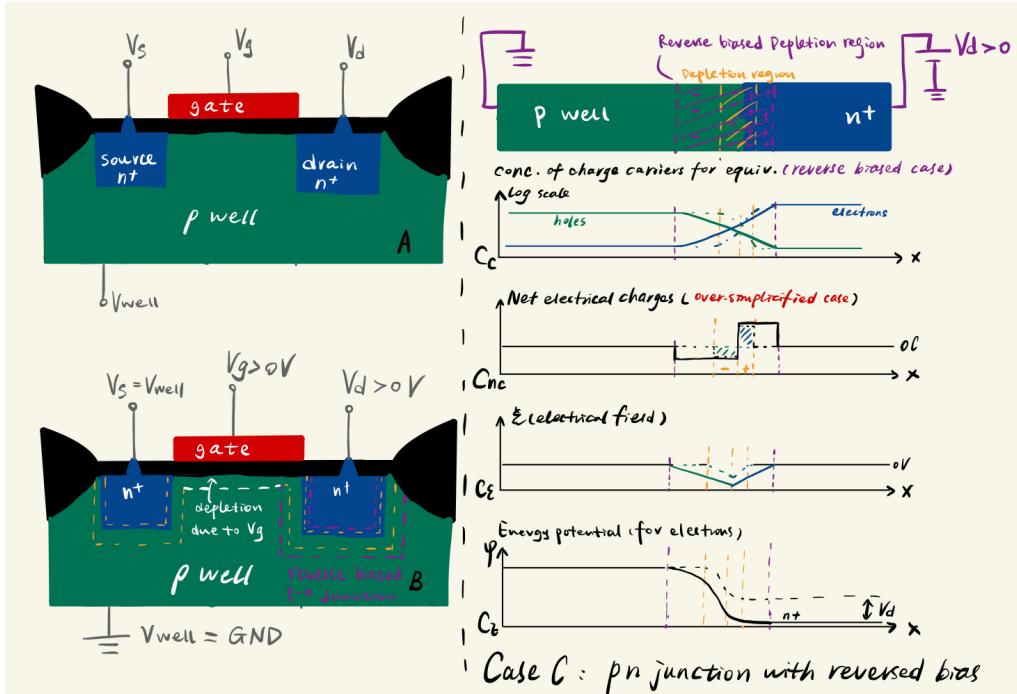


Figure 29: A illustrates how a nFET transistor is typically used by providing four voltage. B shows what are typical voltages applied for neuromorphic usage (note: for sub-threshold, $V_{gs} < V_{thr}$). Case C compares how depletion region properties changes when the p-n junction is reverse biased without considering the effect of V_g in solid line with case B in dashed line. Figure drawn by Wencan Huang

Figure 29 demonstrates how nFET transistors are connected to different voltage in general and illustrates how depletion region changes upon voltage difference acting on drain or source. To understand step by step, case C deliberately ignored the effect introduced by V_g through a simplification assuming that $V_g = 0V$. In this case, the p-n junction between highly n doped drain and p well is reverse biased, which expands the depletion region. While since the p-n junction between source and p well has no voltage difference, its situation is the same as in case B.

The properties changed for p-n junction between drain and p well is shown in the four plots, and one important information is that the energy potential for electron in drain is dropped exactly by the supplied voltage V_d . And this is because that it is this supplied voltage expands depletion region when there is no external voltage applied, and thus the potential change for electron should be exactly the same value. And recalling the potential gap ϕ_0 introduced in figure 28, the energy potential gap between drain and p well now becomes $\phi_0 + V_d$.

nFET transistor connecting to voltage supply considering V_g

To discuss the effect due to gate voltage, it is necessary to understand the effective capacitance due to depletion region. More mathematics can refer to later sections. [add the hyperlink to related sections here and ensure the symbols are consistent.](#)

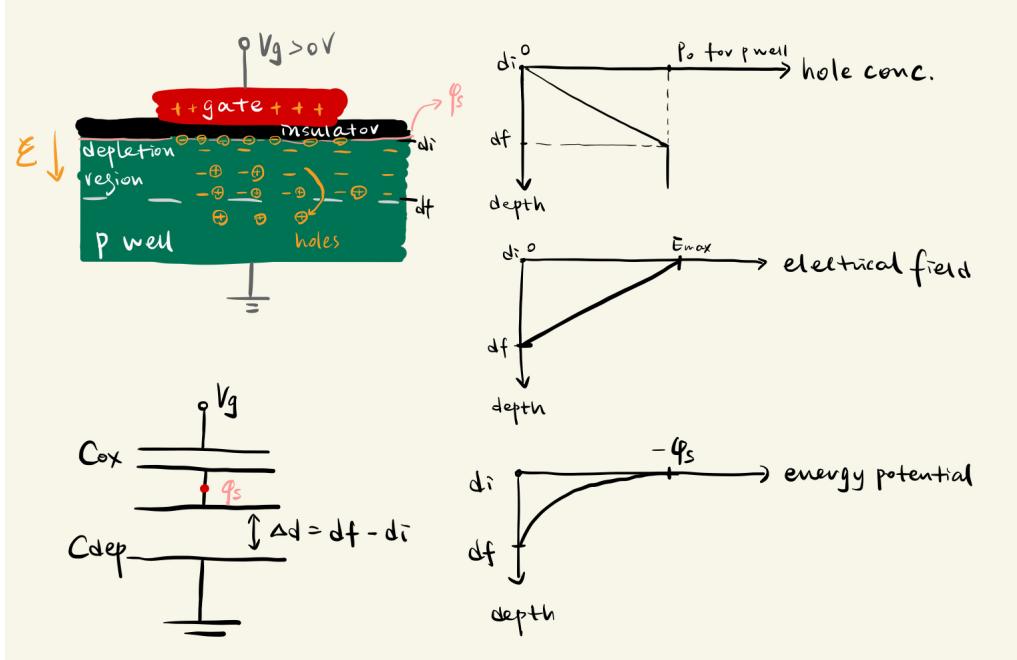


Figure 30: A illustrates how a nFET transistor is affected by V_g as a capacitor. The hole concentration in the first plot changes linearly in log scale according to diffusion - drift balance just as in figure 28. And the energy potential for electron in the third plot is along the negative direction with ϕ_s defined as positive. Figure drawn by Wencan Huang

Along the direction of vertical section, from gate to p well as shown in figure 29 B without considering drain and source, nFET transistor in this case can be considered equivalent to two serial capacitor - one constant capacitor due to the insulator layer, and one causal capacitor due to V_g .

As is shown in three plot in figure 30, the hole concentration in p well can be considered to change exactly the same as in p-n junction in figure 28. With the same simplification as in case A, we can consider the charge density within the depletion region are exactly the same. As a result, this causes a gradual linear change in electrical field through the depletion region, which in terms gives a parabolic change in energy potential, which gives $\phi_s \propto \Delta d^2$, i.e., $\Delta d \propto \sqrt{\phi_s}$.

However, currently we use the term ϕ_s without identifying what this term is. As is demonstrated by the figure of two capacitor, ϕ_s , surface potential refers to the potential for electrons just below the insulator, and its value is mutually determined by V_g and C_{dep} which is caused by V_g .

And using the model of two capacitor without considering the change in Δd and C_{dep} , the change in V_g and ϕ_s should satisfy: $C_{ox}(V_g - \Delta\phi_s) = C_{dep}\Delta\phi_s$, which gives $\frac{d\phi_s}{dV_g} = \frac{C_{ox}}{C_{ox} + C_{dep}}$. Since when $V_g = OV$, $\phi_s = OV$, and for subthreshold region ΔV_g is small, we can assume that $\phi_s = \kappa V_g$ with $\kappa := \frac{C_{ox}}{C_{ox} + C_{dep}}$.

In the above paragraphs, we consider how ϕ_s changes without considering changes in C_{dep} , and this only holds when C_{dep} changes negligibly with respect to changes in V_g or C_{ox} is much larger. Luckily, such assumption holds true and the reason will not be discussed here.

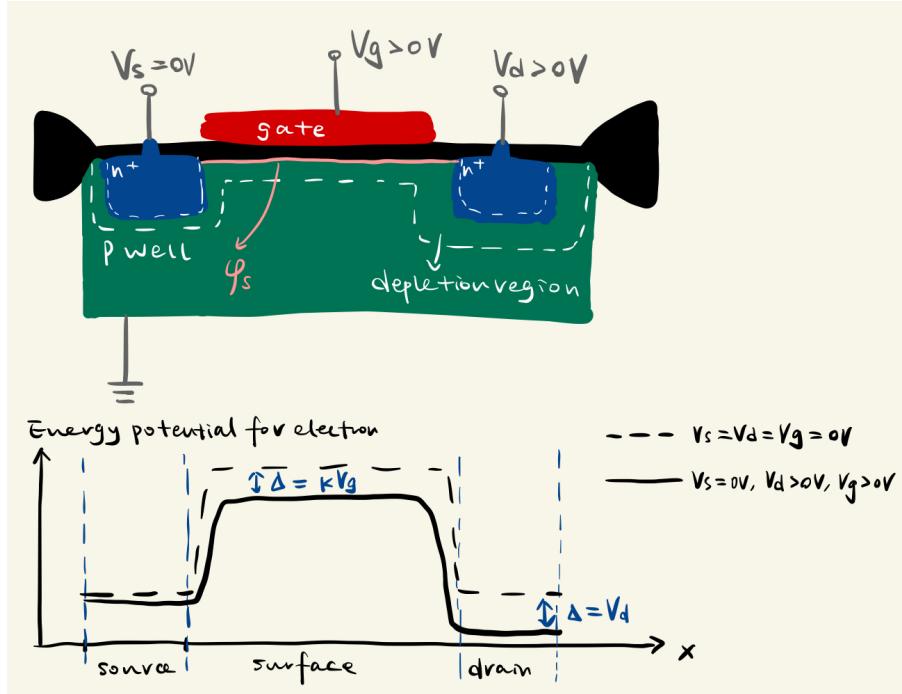


Figure 31: Energy potential for electrons in nFET and the potential is defined by energy/q. Figure drawn by Wencan Huang

As is shown in the plot in figure 31, applying external voltage causes different potential drop in drain side and surface potential as explained above.

Here we need to introduce a fact that the probability of an electron at a specific energy state in solid state material is approximated as Fermi-Dirac distribution: $F(E) = (1 + e^{(E-E_F)/kT})^{-1}$, where E_F is a constant related to material. In general, the term $E - E_F$ is large enough so that the distribution can be simplified as Boltzmann distribution $F(E) = e^{-(E-E_F)/kT}$, and this is where the exponential term in subthreshold current comes from.

Recall that the subthreshold current is represented as $I_{ds} = I_0 \cdot e^{\kappa V_g / V_T} \cdot (e^{-V_d / V_T} - e^{-V_s / V_T})$ (remember, electron flow is in the opposite direction of current), and link this to section 3.1.1, we can realize that V_g affects the general change through affecting ϕ_s and then together with V_d and V_s , the current is determined.

3.4 Subthreshold Operation

In section 3.2, you were introduced to the structure of both n- and p-type MOSFETs. You were also introduced to the functionality of the MIS capacitor in the previous chapter. Let's finally have a look how to *actually* generate a current with a transistor. Remember from section 2.3, that our MOS capacitor has four different operation modes: accumulation, flatband, depletion and inversion. The modes we are interested in are depletion and inversion.

3.4.1 Prelude: Drift and Diffusion Current

The current that is generated in depletion mode is caused by diffusion and we say that we operate in the subthreshold or weak inversion regime. On the other hand, the current generated in inversion mode is caused by drift and we say that we operate in the superthreshold or strong inversion regime. Let's recap what exactly drift and diffusion currents are.

A drift current is caused by an applied electrical field. The field's electrical force defines the strength and direction into which charged particles are pulled. As equally charged particles repel each other and are attracted to opposite charges, negatively charged particles are pulled towards

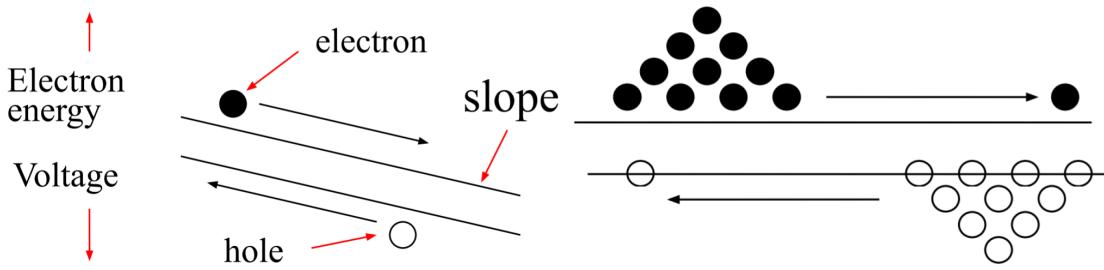


Figure 32: The concept of drift (left) and diffusion (right) current. Adapted from Lecture Notes.

the more positive side of the field and vice versa. Due to the generated movement of charges, we get a current. This current is defined by the following equation:

$$I = qn\mu\epsilon \quad (14)$$

where q is the electron charge, n the carrier density, μ the carrier mobility and ϵ the electrical field. This mechanism is also visualized in Drift subfigure. An electron is pulled towards a point with lower electron energy, i.e. voltage. On the other hand, a hole is pulled towards a point with higher voltage. The slope corresponds to the electrical field ϵ .

A diffusion current is caused by a difference in concentration. We experience diffusion on a daily basis, for example when dipping our tea bag into hot water and seeing it spread or when watching the smoke of a nearby fire that diffuses into the air. Diffusion describes the movement from a region of higher concentration to a region of lower concentration. In our case, this movement of charged particles generates a current which can be described as follows:

$$I = -qD \frac{dN}{dz} \quad (15)$$

where q is the electron charge, D the diffusion constant and $\frac{dN}{dz}$ the concentration gradient. The higher the concentration difference, the higher the resulting current. The concepts of diffusion is also visualized in the diffusion subfigure.

It is important to note that in both regimes diffusion **and** drift current occur. However, in subthreshold the electric field is so weak that it can be neglected next to the diffusion current and vice versa. As you have seen previously, the mode we operate in depends on our gate voltage and we switch from the sub- to the superthreshold regime once our gate voltage crosses a threshold voltage at which point electrons become the majority carriers assuming a p-type channel.

3.4.2 Let's start deriving equations

Let's have a look at our MOSFET when we don't apply any gate voltage. For the following derivations we assume an n-type MOSFET. Remember that an nFET consists of a p-type body and an n-type source and drain. Figure 33 visualizes an nFET transistor and its conduction band E_c . At the contact points between the n-type source and drain with the p-type body, we have a pn-junction. As introduced in section 2.2, the conduction band of the p-type has a higher electron energy than the conduction band of the n-type. This energy difference creates a barrier between the source and the drain that prevents any current from flowing. We denote this energy difference the built-in potential Θ_0 . It is the energy an electron requires to move from the n- to the p-type.

What happens to our energy barrier as soon as we increase the gate voltage? The positive charge at the gate repels the free holes in the p-substrate and a region with only fixed negatively

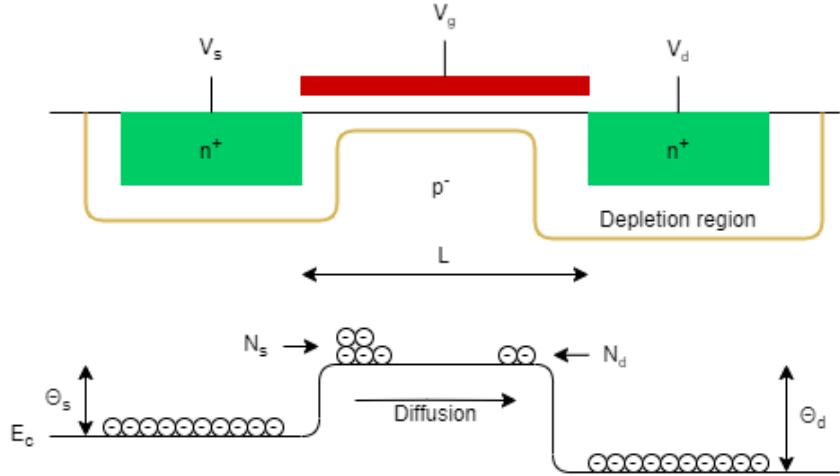


Figure 33: nFET MOSFET structure and its corresponding conduction band.

charged ions, the depletion region, remains. As introduced in section 2.3, the charge of this region is defined by its surface potential ψ_s . By adding an external voltage to the p-type, we lower its conduction band and therefore decrease our energy barrier. On the other side, remember that our source and drain are also connected to an external voltage V_s and V_d . These allow us to change the conduction band of the n-type source and drain, and consequently the energy barrier, as well. We can therefore describe the height of the energy barriers at the source and the drain as follows:

$$\Theta_s = \Theta_0 - q\psi_s + qV_s = \Theta_0 - q(\psi_s - V_s) \quad (16)$$

$$\Theta_d = \Theta_0 - q\psi_s + qV_d = \Theta_0 - q(\psi_s - V_d) \quad (17)$$

We assume that the surface potential ψ_s along the channel is constant. Note that $V_s, V_d \geq 0$ in order to keep our transistor reverse biased, so we cannot simply add a negative voltage to our source and drain to decrease the energy barrier. The height of the energy barrier determines how many electrons can diffuse from the n-type to the p-type channel. The electron densities at the edges of the p-type channel can be described by the following equation:

$$N_s = N_0 e^{-\Theta_s/kT} \quad (18)$$

$$N_d = N_0 e^{-\Theta_d/kT} \quad (19)$$

As expected, the carrier density increases exponentially for a decreasing energy barrier. It should be clear now why we need $V_d > V_s$ for a current to flow. If $V_d = V_s$, $N_d = N_s$ and there is no change in carrier concentration across the channel and hence no diffusion current. For increasing values of V_d , we increase the energy barrier on the drain side of the channel which leads to smaller concentration of electrons. The concentration gradient between the source and the drain can simply be described by:

$$\frac{N}{dz} = \frac{N_s - N_d}{L} \quad (20)$$

where L is the length of the channel. In a semiconductor, the diffusion current is given by:

$$I = -qWtD_n \frac{dN}{dz} \quad (21)$$

where D_n is the diffusion constant, W the width of the channel and t the depth of the channel. By inserting the retrieved formula for our concentration gradient, we finally get the equation for our subthreshold current:

$$I_{ds} = -q \frac{W}{L} t D_n (N_d - N_s) = -q \frac{W}{L} t D_n e^{\frac{\psi_s}{U_T}} (e^{\frac{-V_d}{U_T}} - e^{\frac{-V_s}{U_T}}) = I_0 e^{\frac{\psi_s}{U_T}} (e^{\frac{-V_s}{U_T}} - e^{\frac{-V_d}{U_T}}) \quad (22)$$

where $I_0 = q \frac{W}{L} t D_n N_0 e^{\frac{-\psi_0}{U_T}}$ is the so-called leakage current. Leakage current is current that is always there, despite you not wanting it to be there - it's a physical reality as you always have electrons going around. It can be easily extracted in an experimental setup by setting the voltage difference between the gate and the source to zero, i.e. $V_{gs} = 0$, and measuring the remaining current. The problem with (22) is that it depends on the surface potential ψ_s that we cannot directly control. However, as introduced in section 2.3, ψ_s is related to the gate voltage V_g via:

$$\frac{\Delta \Psi_s}{\Delta V_g} = \frac{C_{ox}}{C_{ox} + C_{dep}} := \kappa \quad (23)$$

Substituting this relationship into (22) yields the important equation:

$$I_{ds} = I_0 e^{\frac{\kappa V_g}{U_T}} (e^{\frac{-V_s}{U_T}} - e^{\frac{-V_d}{U_T}}) \quad (24)$$

We denote the current I_{ds} as the current flow from the drain to the source. Note that current flows **in the opposite direction** of the electrons. When rewriting (24), we see that the subthreshold current I_{ds} actually consists of two components: the current from the drain to the source of the transistor, the so-called forward current, minus the current from the source to the drain of the transistor, the so-called reverse current.

$$I_{ds} = I_0 e^{\frac{\kappa V_g - V_s}{U_T}} - I_0 e^{\frac{\kappa V_g - V_d}{U_T}} = I_f - I_r \quad (25)$$

We can further rewrite the current equation as follows:

$$I_{ds} = I_0 e^{\frac{\kappa V_g - V_s}{U_T}} (1 - e^{\frac{-V_{ds}}{U_T}}) \quad (26)$$

where V_{ds} is the voltage difference between the drain and the source $V_d - V_s$. We can see that for large values of V_{ds} , the reverse current of the equation eventually becomes negligible. When this happens, we are said to operate in the saturation region. The regime in which both the forward and the reverse current shape I_{ds} is called the triode, linear or ohmic region. The point at which we move from the linear to the saturation regime is approximately at $V_{ds} = 4U_T \approx 100mV$. The relationship between V_{ds} and I_{ds} is also visualized in figure 34 for different values of V_{gs} . Note that all values of V_{gs} are in subthreshold and it only influences the strength of the resulting current.

Let's sum up the equations we derived for an nFET transistor in the subthreshold regime.

Subthreshold nFET I_{ds} current

- Triode/ Linear/ Ohmic Region

$$I_{ds} = I_0 e^{\frac{\kappa V_g - V_s}{U_T}} (1 - e^{\frac{-V_{ds}}{U_T}}) = I_f - I_r \quad (27)$$

- Saturation Region

$$I_{ds} = I_0 e^{\frac{\kappa V_g - V_s}{U_T}} = I_f \quad (28)$$

In the saturation region the reverse current is so small that we typically omit it. This mathematically makes sense as the V_{ds} component is a lot smaller than the V_{gs} component, thus vanishing in the equation. Typically, we start considering that we are in saturation starting from $V_{ds} > 4U_T$. This is very important!

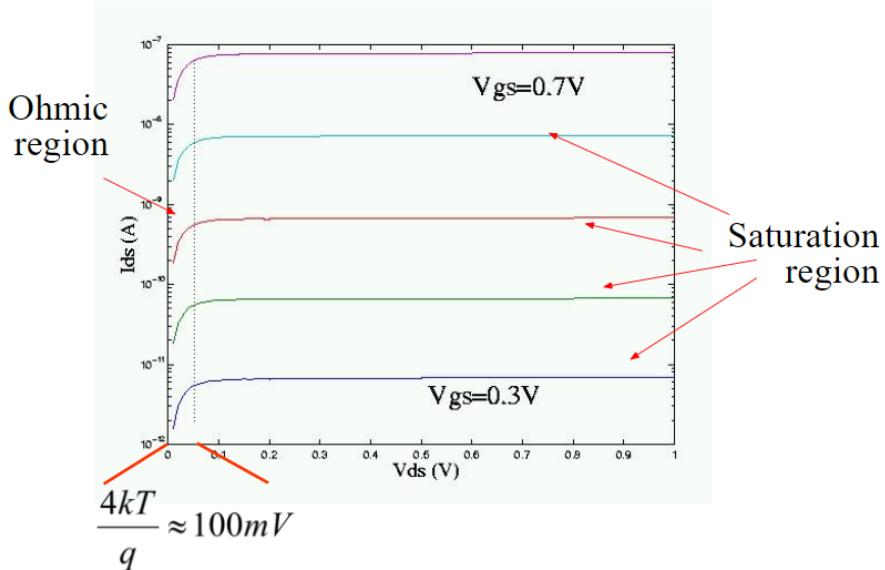


Figure 34: Relationship between the drain-to-source voltage V_{ds} and the current I_{ds} .

Note in the very important figure 34 the following things:

- Current grows linearly in the Ohmic region
- Current is pretty much constant past the $V_{ds} > 4U_T$ (this is not exactly true, as we will see in a later part with the Early Effect).
- The shape of the I_{ds} vs V_{ds} is the same when we have different gate voltages - **as long as we keep in Subthreshold!!!!**

3.5 Superthreshold Operation

When we continue to increase the gate voltage, we attract more and more free electrons to the surface of the channel. Once they become the majority carriers in the channel, the channel becomes n-type and is said to be **inverted**. We operate in the *superthreshold*, or *above-threshold* regime. In the superthreshold regime, the electrical field becomes so strong that it is the main cause for the current flow and makes the previously calculated diffusion current negligible. Remember that the drift current caused by an electrical field is determined by the following equation:

$$I = qn\mu\epsilon Wt \quad (29)$$

where q is the electron charge, n the carrier concentration, μ the mobility of electrons, ϵ the electrical field and W and t the width and the depth of the channel respectively. In the previous section, we have seen that the charge Q_i that accumulates in the inversion region of our channel is dependent on the channel's MOS structure.

$$Q_i = C_{ox}(V_g - V_T) \quad (30)$$

We can express our charge concentration qn in terms of Q_i .

$$qn = \frac{Q_i}{t} \quad (31)$$

Note that Q_i is the inversion charge per unit area. For simplicity we assume that our electrical field is constant. We can therefore simply express ϵ by the voltage difference across the channel.

$$\epsilon = \frac{V_d - V_s}{L} \quad (32)$$

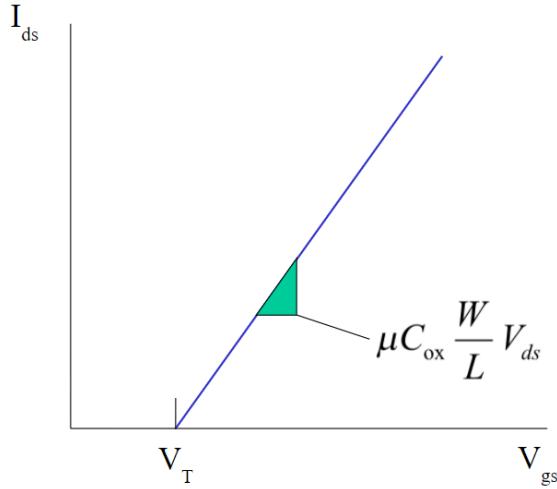


Figure 35: Relationship between the gate voltage V_{gs} and the current I_{ds} for small values of V_{ds} .

where L is the length of the channel. Substituting these relationships into our equation for the drift current (29) yields our current equation in superthreshold.

$$I = C_{ox}(V_g - V_T)\mu \frac{V_d - V_s}{L} W = \beta(V_g - V_T)(V_d - V_s) \quad (33)$$

where $\beta = \mu C_{ox} \frac{W}{L}$. Assuming that β and V_{ds} are constant, we can see that there is a linear relationship between the current I_{ds} and the gate voltage V_{gs} . This relationship is also demonstrated in figure 35.

Effective Threshold Voltage

In order to switch between the sub- and the superthreshold regime, we have to increase our gate voltage until it crosses a specific voltage value, the threshold voltage V_T . So far, we have assumed that V_T is a fixed value, however in reality this is not the case. When we increase the source voltage, hence decreasing the voltage difference V_{ds} , our current decreases as a consequence of the increased energy barrier between the source and the channel. How much do we have to change our gate voltage to retrieve our original current? An increase of V_s can be counter balanced by an increase of ϕ_s . For $\Delta V_s = \Delta \phi_s$, our energy barrier, and hence the generated current, remains the same. Remember that $\frac{\Delta \phi_s}{\Delta V_G} = \kappa$, so we have to increase our gate voltage by a factor of $\frac{1}{\kappa}$ more than the source voltage. But what does this have to do with the threshold voltage? At the threshold point, electrons become the majority carriers in the channel. These mobile electrons are mainly attracted from the source. For increasing V_s , we therefore have to increase our gate voltage by the factor of $\frac{1}{\kappa}$ more in order to attract enough free carriers to invert our channel. The actual, **effective**, threshold voltage is hence dependent on the source voltage.

$$V_T = V_{T0} + \frac{V_s}{\kappa} \quad (34)$$

Let's have a look at our channel again. We have previously assumed that the inversion charge Q_i is constant along the channel. In reality, it decreases from its highest concentration at the source end to its lowest concentration at the drain end. The charge is dependent on the threshold voltage at the channel ends and consequently on the source and drain voltages as given by (34). For the charge at the source and the drain end we get:

$$Q_s = C_{ox}(V_g - V_T) = C_{ox}(V_g - V_{T0} - \frac{V_s}{\kappa}) \quad (35)$$

$$Q_d = C_{ox}(V_g - V_T) = C_{ox}(V_g - V_{T0} - \frac{V_d}{\kappa}) \quad (36)$$

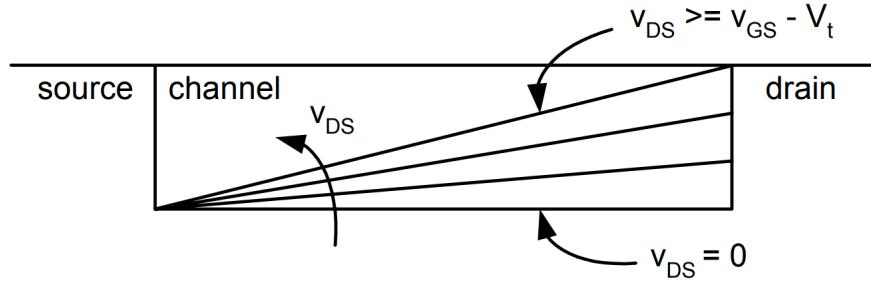


Figure 36: Different channel shapes for different values of V_{ds} [Taken from http://in.ncu.edu.tw/ncume_ee/harvard-es154/lect_12_MOSFETs.pdf].

So for a constant gate voltage, the charge decreases for an increasing V_s and V_d . As $V_d > V_s$, the charge concentration at the drain end is indeed lower just as expected. If we continue to increase V_d , the charge at the drain end of the channel Q_d will decrease until it eventually becomes zero. The channel is said to be "pinched off". Figure 36 demonstrates how the channel shape changes for different values of V_{ds} . The so-called pinchoff point is the point in the channel where the inversion charge appears first. The electrons are moved from the drain end into the drain region by the electrical field between the drain and the channel. The drain voltage, however, does not influence the current, i.e. the electron movement from the source along the channel, anymore and I_{ds} remains constant for increasing values of V_{ds} . Similar to the subthreshold regime, we now say that the current is saturated and that we operate in the saturation region. This happens once the voltage difference V_{ds} corresponds to the overdrive voltage V_{ov} , so when the charge difference equals the charge concentration at the source. For simplicity, we assume a fixed threshold value again and get the following saturation voltage:

$$V_{ds,sat} = V_{gs} - V_T = V_{ov} \quad (37)$$

By replacing V_{ds} with $V_{ds,sat}$, we can define the saturated current $I_{ds,sat}$.

$$\frac{1}{2}\beta(V_{gs} - V_T)^2 = \frac{1}{2}\beta V_{ov}^2 \quad (38)$$

The additional factor $\frac{1}{2}$ can be explained by a more detailed derivation of the drain current I_{ds} which does not assume a constant electrical field along the channel. Please refer to the Textbook if you want to find out more. Figure 37 visualizes the relationship between V_{ds} and I_{ds} for different gate voltage V_{gs} .

Special current I_s

Finally, let's define a specific current you should know about: the special current I_s . It defines the border between weak and strong inversion and is particularly useful during designing. It approximately corresponds to the current at the overdrive voltage $V_{ov} = U_T$.

$$I_s = 2\mu C_{ox} \frac{W}{L} U_T^2 = 2\beta U_T^2 \quad (39)$$

Let's sum up the equations we derived for an nFET transistor in superthreshold regime.

Superthreshold nFET I_{ds} current

- Triode/ Linear/ Ohmic Region

$$I_{ds} = \beta(V_{gs} - V_T)V_{ds} = \beta V_{ov} V_{ds} \quad (40)$$

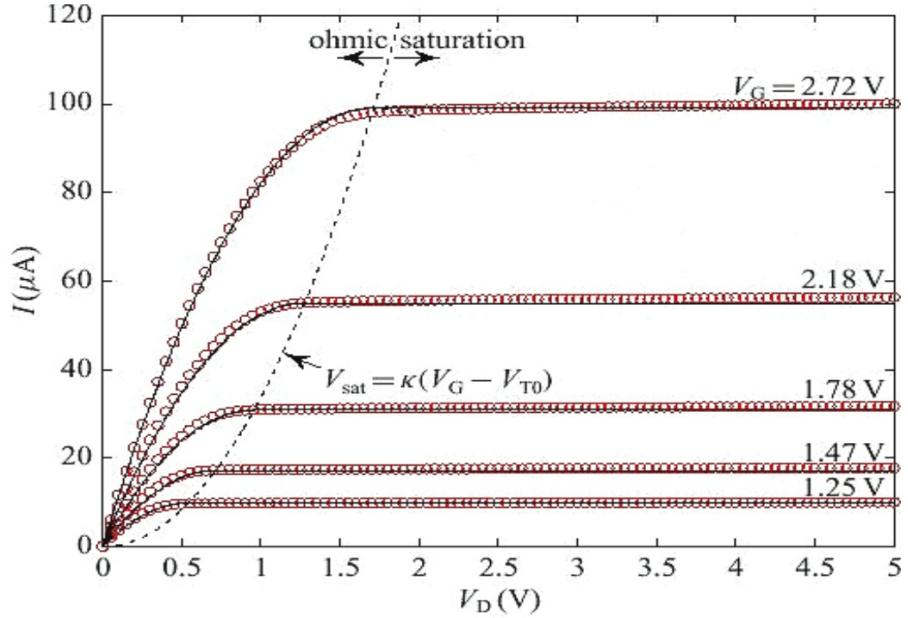


Figure 37: Relationship between V_{ds} and I_{ds} for different gate voltage V_{gs} .

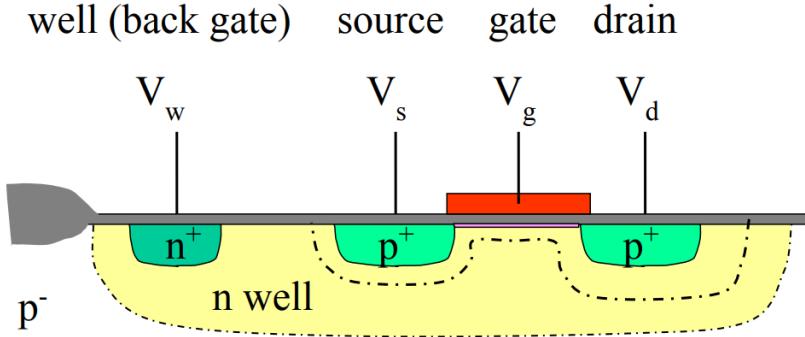


Figure 38: Structure of a pFET MOSFET.

- Saturation Region

$$I_{ds} = \frac{1}{2} \beta (V_{gs} - V_T)^2 = \frac{1}{2} \beta V_{ov}^2 \quad (41)$$

3.6 pFET MOSFET

In the previous two sections we introduced how an nFET MOSFET transistor operates in both the sub- and superthreshold regime. As mentioned in chapter 2, there exists another type of MOSFETs: the pFET transistor. Remember that in a pFET, the doped substrates are inverted. The source and the drain contain free holes as the majority carriers and are therefore of p-type. Both lie within a body with electrons as the majority carriers, the n-well. This n-well is connected to the highest voltage V_{dd} . The structure of the pFET is shown in figure 38.

The pFET transistor behaves exactly like the nFET transistor but inverted. What exactly does that mean? The transistor is turned off when the gate voltage V_g is at the highest voltage V_{dd} . As there is no potential difference between the gate and the n-well, there is no surface potential and hence no current along the channel. When we now decrease the gate voltage of the

pFET, the negative potential difference repels the free electrons within the n-well at the surface of the channel and we get a region of fixed positive ions only, the depletion region. As described in the previous section, the diffusion in this region creates a current I_{sd} . If we further decrease the gate voltage, free holes eventually become attracted to the channel surface and an inversion region of p-type is created. Similar to the previous derivations, we get the following equations for the current in a pFET MOSFET. Note that they are exactly like the nFET equations but with an inverted sign.

Superthreshold pFET I_{sd} current

- Triode/ Linear/ Ohmic Region

$$I_{sd} = \beta(V_{sg} - |V_T|)V_{sd} \quad (42)$$

- Saturation Region

$$I_{sd} = \frac{1}{2}\beta(V_{sg} - |V_T|)^2 \quad (43)$$

Subthreshold pFET I_{ds} current

- Triode/ Linear/ Ohmic Region

$$I_{sd} = I_0 e^{\frac{-\kappa V_g + V_s}{U_T}} (1 - e^{\frac{V_{ds}}{U_T}}) \quad (44)$$

- Saturation Region

$$I_{sd} = I_0 e^{\frac{-\kappa V_g + V_s}{U_T}} \quad (45)$$

3.7 Bulks, wells and biasing the MOSFET Bulks

Good question. A lot is actually going on, but let's try to keep it short and simple. Remember from chapter 0 (which I hope you reviewed!) that voltage is all about potential *difference*, so you need "point of references" that are somewhat common to everywhere in your circuit. It so happens that in a circuit, all transistors do not necessarily share the same bulk, which is typically the natural reference potential of MOSFETs. We therefore need a common reference potential that multiple transistors can have access to: this is called the common bulk potential. There are two common bulk potentials in a circuit: the lowest (ground) is denoted V_{ss} and the highest is denoted V_{dd} . You never want charge carriers from the source/drain of transistor to go into the bulks (except in very rare scenarios not considered in this course) - to avoid that, source and drain diodes are **reverse-biased** to the bulks. To do so, in an nFET, the bulk is typically at V_{ss} and source will always be slightly higher (or equal) potential: $V_s \geq V_{ss}$. This ensures the reverse bias as we apply a higher potential to the n-doped semi conductor than the p-doped semi conductor it is in contact with. Same holds in PFET, but in reverse. All voltages in the nFET are referenced to V_{ss} , which is typically the 0 V and lowest potential (ground) of nFETs - so all voltages we'll be dealing with are typically positive. In the pFET, all voltages are referenced to V_{dd} . Connections to V_{ss} are marked as connection to ground, and connections to V_{dd} with a slanting line.

What does that mean for our derived equations? You guessed right, we have yet made another assumption and omitted our reference bulk voltages. What we are in reality interested in is how much we change our voltages V_g , V_s and V_d **compared to** our bulk voltages. Let's have a look at our pFET again. Remember that in a PFET our n-well is biased to the highest voltage V_{dd} and our transistor is off when our gate voltage is at V_{dd} as well. In this case, there is no potential difference between the gate and the n-type channel and therefore our doping concentration equals our mobile carrier concentration. No current is flowing. When we now decrease the gate voltage, a depletion and eventually an inversion region is created across the channel. This is dependent on the potential difference between our gate and channel, i.e. how much our gate voltage now differs **from** our bulk voltage. Let's rewrite the subthreshold equations to incorporate the difference we are actually interested in:

- For the subthreshold NFET in saturation:

$$I = I_0 e^{\frac{\kappa(V_g - V_{ss}) - (V_s - V_{ss})}{U_T}} \quad (46)$$

- For the subthreshold PFET in saturation:

$$I = I_0 e^{\frac{-\kappa(V_{dd} - V_g) + (V_{dd} - V_s)}{U_T}} \quad (47)$$

Great, so now you know how the bulk voltage modifies our generated current. However, we usually omit V_{dd} and V_{ss} in our equations for simplicity but it is important to know that they are our reference points. As you will see later on in the section on the pFET source follower, in some cases it actually might be beneficial to set them to a value that differs from V_{dd} or V_{ss} .

3.8 Transistor Conductance

Transistors, like any piece of electronics, have some kind of inherent conductance. In Transistors, conductance corresponds to the differential change in current due to differential change in each of the terminal voltages. We there can distinguish between different types of conductances. The source, drain and gate conductance can be defined, both in subthreshold and in superthreshold regimes. We will here only focus on Subthreshold regime and without getting into derivation details which are just the derivatives, we reach:

- Gate conductance:

$$g_{mg} = \frac{\partial I}{\partial V_g} = \frac{\kappa I}{U_T} \quad (48)$$

- Source Conductance

$$g_{ms} = \frac{\partial I}{\partial V_s} = \frac{I_0 e^{\kappa V_g - V_s} / U_T}{U_T} \quad (49)$$

- Drain Conductance

$$g_{md} = \frac{\partial I}{\partial V_d} = \frac{I_0 e^{\kappa V_g - V_d} / U_T}{U_T} + \frac{1}{V_E} \quad (50)$$

We will look at Early Voltage V_E in an upcoming section.

There are also other type of conductances, which we will look at when reaching circuit. Mainly, there is the transconductance, which relates output current to input voltage $g_m = \frac{\partial I}{\partial V_{in}}$. There is also the output conductance $g_{ds} = \frac{\partial I}{\partial V_{ds}}$. We'll look at these in dedicated sections. All you should remember for now is that transistors have conductances which relate the differential change in current produced with the differential change in various type of voltage applied.

3.9 Second Order Effects

In the derivation of the current-voltage characteristics of the MOSFET, we assumed that certain properties are constant under all operating conditions. These ideal assumptions do not apply in a real device. Some examples of the non-idealities in a MOSFET are described here.

3.9.1 Prelude: How transistor width and length impact operation

Remember that both the sub- and the superthreshold current are scaled by a parameterized factor: I_0 and β where

$$I_0 = q \frac{W}{L} t D_n N_0 e^{\frac{-\phi_0}{U_T}} \quad (51)$$

$$\beta = \mu C_{ox} \frac{W}{L} \quad (52)$$

Both factors are dependent on the ratio between the transistor's width and its length $\frac{W}{L}$. This is an important condition to consider when designing MOSFETs. Simply looking at this condition, it seems desirable to keep our transistor length as small as possible in order to get a power efficient transistor. In reality, the length of the channel defines the impact of second order effects onto our generated current flow and it is typically desired to have a longer channel. An ideal width to length ratio is therefore a very tricky design decision. The following sections will introduce the most important second order effects and in particular how they are shaped by the transistor's length.

3.9.2 Transistors in real life, and the problem of mismatch

You may have heard of Moore's law³⁷. The observation is named after Gordon Moore, the co-founder of Fairchild Semiconductor and Intel (and former CEO of the latter), who in 1965 posited a doubling every year in the number of components per integrated circuit. The fundamental component of the integrated circuit is the transistor - so this implies that transistors are supposed to halve in size every year! Due to extremely fancy industrial processes, the smallest transistors today are as small as a few nanometers length (length from source to drain). We're now reaching the physical limits of Moore's law as the transistor has a theoretical minimum length, and it wouldn't function the same way if we ever manage to make it smaller. Besides this issue, we should note something: transistor function is impacted by their sizes. There are many *second order effects* that are typically omitted in the modelling because they are negligible, however, because we're dealing with the limitation of the physics, these second order effects become important. We'll look at these in more details in the next chapter, but what I'd like you to remember is the fact that a difference in size of 1 mm between two Ikea shelves will not matter when you build your piece of furniture, because 1mm is negligible compared to the overall size of the shelf. However, if you have a 1cm difference, it might start to be a problem. Similarly with the transistor, because we're dealing with extremely small components, and because the industrial processes cannot realistically be precise to the 10^{-12} , you always end up with slight size differences between the transistors within a circuit, which lead to some differences in the way they function, and is the source of many limitations in modern VLSI. This problem is named **transistor mismatch** and it is going to appear very often in your lab plots, as you will see when you start experimenting with the Class Chip.

3.9.3 The Early Effect

Very, very important thing to start with: it's named after James Early, and is not related to being early, late or anything time related. I wanted to mention it first because to this day I unconsciously picture it as something happening before something else. The lecture notes state: "When deriving the I-V (Current-Voltage) characteristics of nFETs in the previous chapter, we assumed that the current is constant when working in saturation ($V_{ds} > 4U_T$). It so happens that this assumption is not sufficient, especially with short length MOSFETs, for which the drain voltage can modulate the channel current, even in saturation." So what is exactly happening here? Let's start by reminding ourselves the (nFET) relationship and the theoretical graph linking drain current I_{ds} to drain to source voltage V_{ds} for fixed V_{gs} :

$$I_{ds} = I_{n0} e^{\frac{\kappa_n V_g}{U_T}} \left(e^{\frac{-V_s}{U_T}} - e^{\frac{-V_d}{U_T}} \right) \quad (53)$$

We also should look at the graphical relationship this relation implies:

³⁷Moore's law was actually a term coined by Carver Mead!

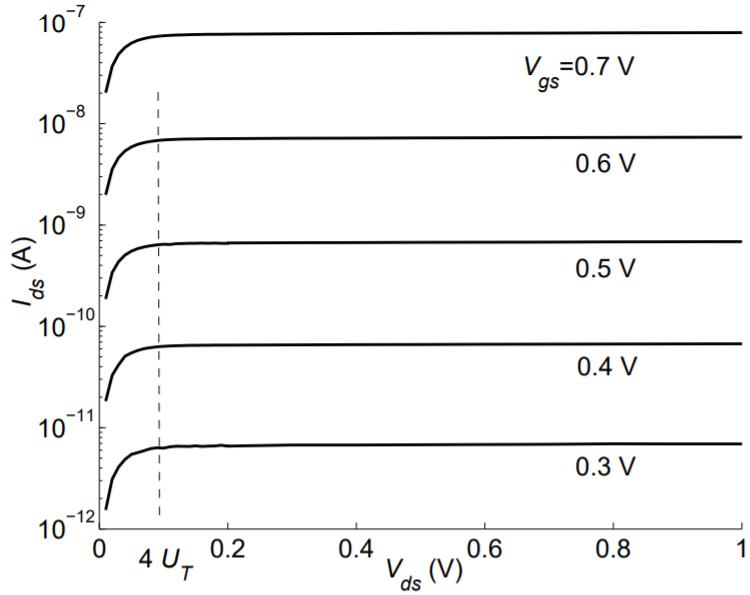


Figure 39: Relationship between V_{ds} and I_{ds} for fixed V_{gs} . Determines the voltage at which we switch from Ohmic to Saturation region. Adapted from Lecture Notes.

One should notice in the equation and that I_{ds} has an almost insignificant dependence on V_{ds} , because of the exponential term, where the term in parenthesis just vanishes to V_s as V_d (and thus V_{ds}) increases. We consider this to start being true when $V_{ds} > 4U_T$, where we go from the "Ohmic" Region to the "Saturation region". This is clearly apparent in figure 39, where changing V_{ds} does not affect I_{ds} past the $4U_T$ threshold - thus yielding a constant relationship between V_{ds} and I_{ds} past the threshold (again, taking V_{gs} as fixed).

Now surprise surprise, this model is not correct in practice. This is called the Early effect, and basically introduces the practical problem of slightly rising I_{ds} when increasing V_{ds} , even past the the saturation threshold. This rate of change is critically impacted by the geometry of the transistor, mainly its W/L ratio. Effectively, when increasing V_d , the *effective length* L_{eff} of the transistor decreases. This is because the pinchoff region extends further along the channel away from the drain (see figure below). This has a particularly important impact when we are dealing with short transistors, as the shift in pinchoff is, relatively to the channel length, more important (see figure 40).

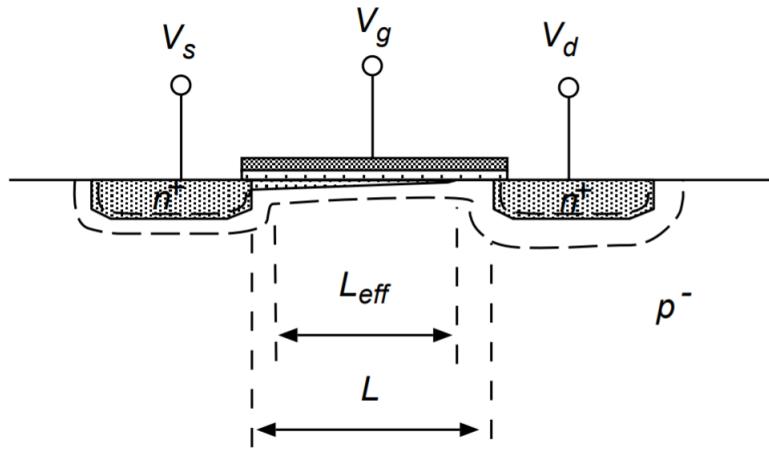


Figure 40: The effective channel length L_{eff} of a transistor operating in the above-threshold saturation region decreases with increasing V_d because the pinchoff point moves into the channel, away from the drain. The effective channel length can be described by the transistor length minus the length of the pinchoff region in the channel. Adapted from Lecture Notes.

Ok great. So what do we do with this information? Well now we need to find a way to quantify this effect and include it in the previous equation to have a more accurate model. For this, we need to use the *Early Voltage* and the following relations:

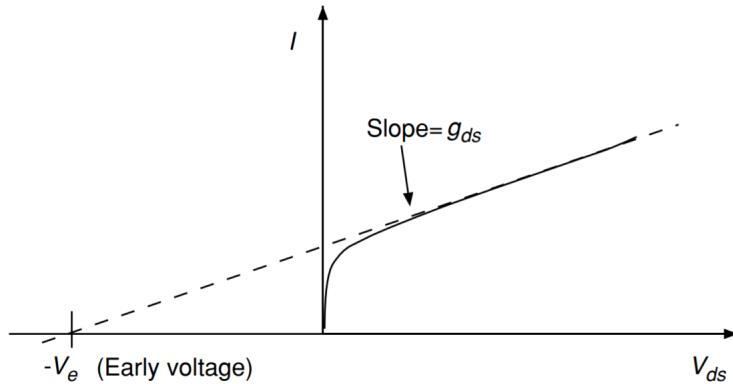


Figure 41: Plot of current versus drain-to-source voltage, showing the slope of the curve g_{ds} in the saturation regime. The intersection of the slope with the V_{ds} axis is called the Early voltage. Adapted from Lecture Notes.

We just established that even in Saturation, there is a slight increase in I_{ds} as a function of increasing V_d - this increase is linear as shown in the figure above, and it has a very specific slope called g_{ds} , also called the *output conductance* of the transistor. It is defined as follows:

$$g_{ds} = \frac{\partial I}{\partial V_{ds}} = \frac{\partial I}{\partial L_{eff}} \frac{\partial L_{eff}}{\partial V_{ds}} = \frac{I}{V_e} \quad (54)$$

V_e is the *Early Voltage*, which is defined as the **absolute value of voltage for which I_{ds} is 0 when in saturation**. It is only a theoretical voltage that allows you to quantify the steepness of the I_{ds} slope in saturation, and thus the extent by which I_{ds} changes as a function of V_{ds} changes. Understanding the precise derivation of the equation requires a substantial amount of device physics, which is out of the scope of these lecture notes. If you would like to understand better, refer to the Textbook, Chapter 3. Now accepting equation 54 as true, we can rewrite our equation for drain current in a more complete manner, as follows:

$$I = I_{sat} + g_{ds}V_{ds} = I_{sat}\left(1 + \frac{V_{ds}}{V_e}\right) \quad (55)$$

Here, $I_{sat} = I_{no}e^{\frac{\kappa_n V_g - V_s}{U_T}}$. Should you really care about this second order effect? Yes, because V_E ranges between 750V and 20V for typical transistors operating in Subthreshold, as shown in the figure below:

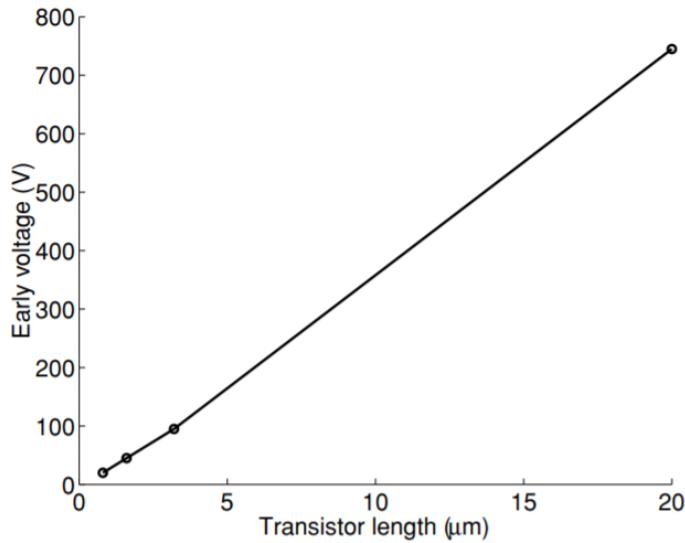


Figure 42: Early voltage versus the transistor length of an nFET fabricated in a $0.8 \mu m$ CMOS process. Adapted from Lecture Notes.

Keep in mind that 20V is a way more significant Early Voltage than 750V, because that equates to a lot steeper slope! This is why Early Voltage is particularly important in shorter MOSFETs: a slight change in V_{ds} will have significant impact on I_{ds} when in Saturation! The smaller the transistor, the smaller V_E , and the smaller V_E , the higher $\frac{\partial I}{\partial V_{ds}}$.

3.9.4 The Body Effect

The *Body Effect*, also called the *Back Gate Effect* relates to the bulk and its influence on transistor operation. In the I-V (Current to Voltage) equations that we have derived so far, the terminal voltages of the transistor are referenced to the bulk. However, the bulk is also an input to the transistor that should be considered in most circumstances, though we often chose to neglect it for simplicity. In the subthreshold region, we can describe the influence of the bulk potential V_b through the series of capacitors C_{ox} and C_d (as we saw also for the gate input). The effect on the surface potential can be written as:

$$\partial\psi_s = (1 - \kappa)\partial V_b \quad (56)$$

In the strong inversion, or above-threshold, regime the influence of the bulk voltage is usually treated as an increase in the threshold voltage of the transistor. If V_b decreases, then there is practically no change in the gate charge because the voltage across the gate oxide is essentially unchanged (the surface potential remains approximately the same). However, the depletion region underneath the gate increases³⁸. Since the negative charge from the depletion region is now larger, less charge is required in the inversion region to balance the gate charge. The inversion region becomes smaller, so leading to a smaller I . To restore I to its original value, we increase the gate voltage V_{gs} . Assuming that we do not forward bias the PN junctions between the drain/source regions

³⁸Remember from the PN Junction that increasing the reverse bias across a PN junction causes the depletion to increase

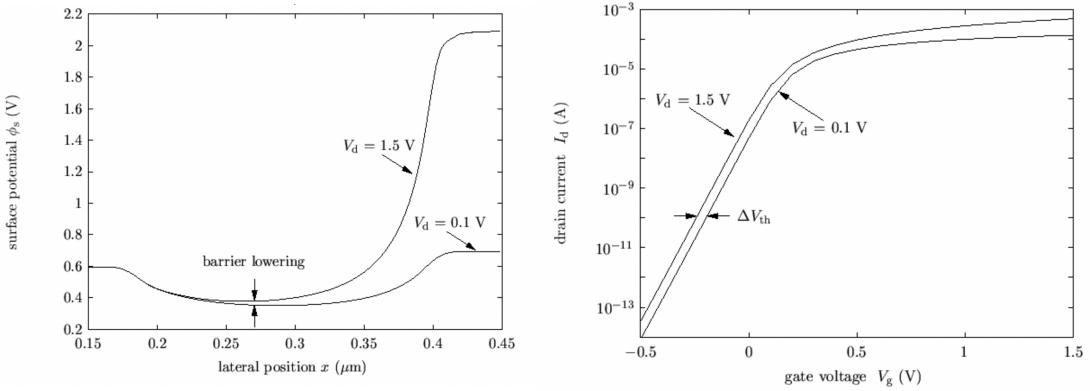


Figure 44: Varying surface potentials (left) and I-V curve (right) for different drain voltages due to Drain Induced Barrier Lowering (DIBL).

and the bulk: What happens if the bulk voltage V_b is increased by ΔV ? This scenario is the same as decreasing V_g , V_s , and V_d by the same ΔV . In the subthreshold region, the change in ψ_s will now be $-\kappa\Delta V$. The barrier height is decreased at both ends of the channel, and the current increases. Hence, the bulk acts like the gate, but it has a weaker influence on the transistor current.

3.9.5 Drain Induced Barrier Lowering (DIBL)

In the below threshold regime there is a potential barrier between the source and the channel region. The height of this barrier is a result of the balance between drift and diffusion current between these two regions. If a high drain voltage is applied, the barrier height can decrease, as indicated in figure 43, leading to an increase in the surface potential ??, which then also leads to an increased drain current as shown in figure ???. The precise derivation of this barrier lowering effect is more complicated and not part of this course, you just have to know, that it exists.

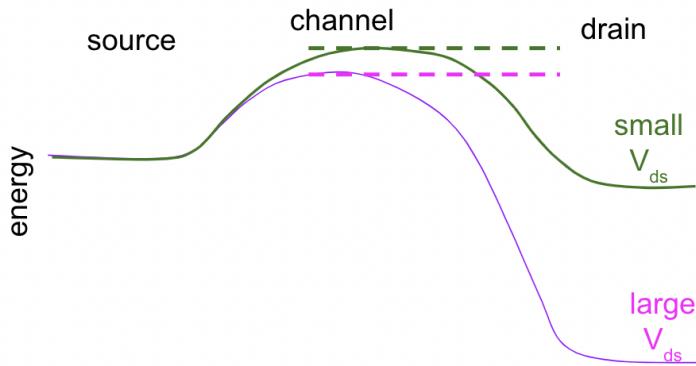


Figure 43: Varying energy barrier for different drain voltages due to Drain Induced Barrier Lowering (DIBL).

To summarize, Drain Induced Barrier Lowering means, that the drain current is controlled not only by the gate voltage, but also by the drain voltage (although rather weakly), because a higher drain voltage decreases the energy barrier between source and channel. It is especially pronounced in small width transistors, as there the drain can influence the source more easily. Since this parasitic effect simply shifts the current up or down, it can be accounted for by a threshold voltage reduction depending on the drain voltage for device modeling purposes. Large drain voltages typically decrease the threshold voltage V_{thr} by $\approx 100mV$.

3.10 Impact Ionization

Finally, large drain voltages can lead to another effect called Impact Ionization. This simply means, that when the drain voltage is very high, it can give the incoming electrons enough kinetic energy, such that, when they bump into other bound electrons (in the valency band), the impact transfers enough energy to free them and elevate them into the conduction band, which leaves a free hole in the valency band, thus creating a new energy-hole pair (figure 46). You can picture this as the drain acting similar to a vacuum cleaner that sucks the electrons in and the higher the voltage, the higher the power of the vacuum cleaner and the faster the electrons get sucked in hence the more kinetic energy they have and if they have enough energy they knock the bound electrons out of their bound state upon impact. The newly created electrons are naturally also affected by the drain voltage and hence also sucked in. As a result, the drain current becomes even larger than the source current (as seen in figure 46).

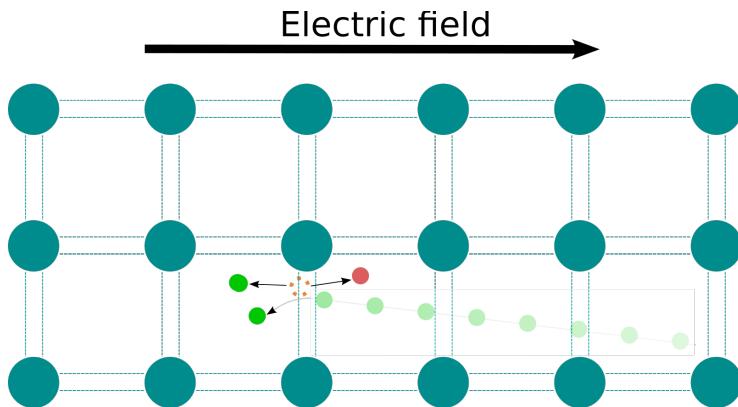


Figure 45

We do not go into detail of the above mentioned drain effects (DIBL and Impact Ionization) but it is important to know that our derived equations neglect several phenomena that occur in real transistors.

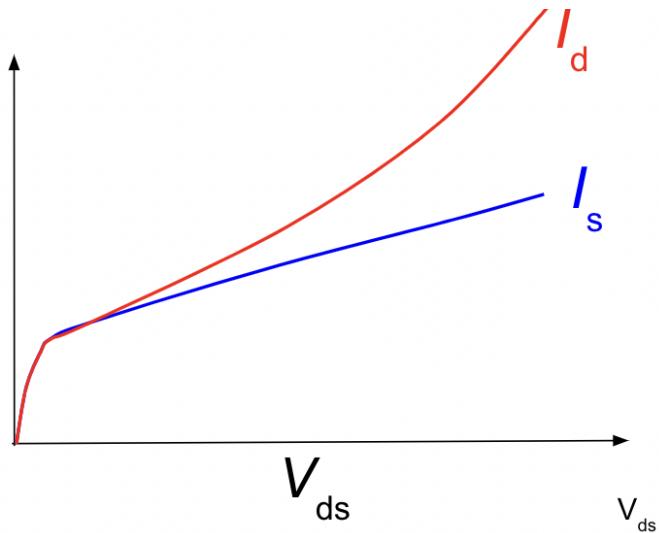


Figure 46

3.11 Concluding thoughts on Subthreshold Regime

In Neuromorphic Engineering 1, we pretty much only focus on Subthreshold equation because they exhibit exponential voltage to current relationship, just like the biological neurons. It also happens that using subthreshold operation is low power compared to above threshold, as the applied gate voltages are significantly lower. However, this comes at a price, because of the exponential $\frac{\partial I}{\partial V_g}$ relationship, we have low stability and noise yields to a lot more problems than above threshold dynamics, which are typically a lot more stable. There is a lot more to say about this, and hopefully I'll have the time to come back to it and add some details.

3.12 Laboratory : Transistors above threshold

As you have understood from reading the theory, current flows through drift and diffusion.

In this section, we will study p-type and n-type transistors when drift is the dominant driver of current.

3.12.1 Voltage threshold and Beta

To find a current dominated by drift, we should look to a gate voltage that is higher than a certain threshold. That's why drift or strong inversion is also described as above threshold operation. It needs to be above (or below in p-fet) a certain threshold and from there we can extract the current driven by drift. This should ring a bell 2.

Source and drain are initialized differently depending whether we are interested in the ohmic or in the saturation region. To study saturation region, the full range of voltage operations is taken, whereas, in the ohmic region, source and drain voltages should differ only slightly (i.e. 0.2 V). This should also ring a bell 2.

We study a range of different gates voltages. Refers to the plot below, to see a real n-FET in ohmic and saturation region.

If you have something similar to this, well done ! You can basically control electricity with your fingers.

The finding are very fundamental here. We proved that with a certain source and drain voltages, and a varying gate voltage, current flows in a transistor following a certain curve. We realized that the equation are only approximation of reality, but they are good approximation of reality.

It has been found that for an n-fet, the voltage gate threshold (the point where the current suddenly increases), was around 0.8V.

Saturation regions start at this index and ohmic regions end at this threshold.

To estimate the real beta from measurement, first, you need to compute the slope of the curve and then apply some transformation to it (the transformation are derived directly from the formulas in ??).

Betas in ohmic region has been found to be a float between 2 and 3, whereas in saturation region they were near 0 .

The betas of a p-fet and an n-fet, the ratio between them should ideally be one. If it's not one, but it's approximately one (0.9), then don't worry, but think : what could be the reason for this discrepancy?

3.12.2 Early voltage

To measure the Early voltage, in n-fet, you need to vary the voltage gate (similarly as before, within a range) and vary the voltage at the drain. This way we can get more data and test whether in reality the Early Voltage is the same with different drain voltages.

You should get something like [Figure].

By fitting a line in the saturation region, you can extract the early voltage. Remember the definition of Early Voltage in ?? .

You will find there is no single Early Voltage, but the measurements you get should be within one order of magnitude. Attention, if you measured a voltage gate of 1.8V than the saturation region will be forced to stay flat by the built in system control, avoid using 1.8V for the gate.

When the voltage gate goes up the saturation region is less flat and the absolute early voltage gets bigger.

3.13 Test Yourself

You should be able to answer the following questions for the exam (mainly taken from the winter study sheet).

- What does it mean for a MOS transistor channel to be accumulated, flat-band, depleted, inverted?
- Knowledge of how subthreshold transistor operation is a diffusion process and why it depends exponentially on the terminal voltages.
- What is the meaning of "saturation"?
- What is the triode or linear operating range?
- I_{ds} vs V_{gs} on log scale.
- Differences between n- and p-fets.
- Typical values of I_0 , κ and subthreshold operating range.
- What are wells and how should the wells be biased relative to the substrate?
- What is the "back gate" or "body effect"?
- How is the back gate related to κ ?
- How to make a MOS capacitor and what is its C-V relationship.
- How transistors work above threshold.
- What is the linear or triode region and what is the saturation region?
- How do they depend on gate and threshold voltage?
- How the Early effect comes about.
- Typical values for Early voltage.
- How to sketch graphs of transistor current vs. gate voltage and drain-source voltage.
- How above-threshold transistors go into saturation and why the saturation voltage is equal to the gate overdrive.
- The above-threshold current equations.
- How above-threshold current depends on C_{ox} and mobility.
- What is DIBL (drain induced barrier lowering) and II (impact ionization)?
- How transconductance and drain resistance combine to generate voltage gain and what is the intrinsic voltage gain of a transistor.
- How transconductance and drain resistance combine to generate voltage gain.

4 Static Circuits

In this chapter, elementary analog VLSI³⁹ circuits are introduced. It is critical to be very comfortable with the dynamics of these circuits to understand the more complex circuits presented in the following chapters. They indeed are the building blocks of many complex circuits which will be encountered later. A few things to remember about the circuits presented here:

- Equivalent circuits are obtained by exchanging MOSFET types (from N-Type to P-Type) and reversing Voltage differences.
- All circuits are derived in steady state. Steady state means that the circuit is in equilibrium where "transient" effects are no longer important.
- Unless explicitly stated, all transistors are assumed to be functioning subthreshold. All these circuits have very different dynamics when working above threshold, which is not the purpose of the course.
- Second order effects, such as the Early Effect, are neglected. So yes, a lot of what is derived here works quite differently in practice!
- For simplicity, MOSFETs are assumed to have a unity width-to-length ratio. Because yes, remember that MOSFET dynamics are affected by their width to length ratio!

Before starting this chapter, one should be fully familiar with the following concepts:

- Electrical Engineering Fundamentals explained in Chapter 0.
- Intuition behind the function of the transistor and how the combination of gate, source and drain voltage yield different current dynamics.
- Basics principle and equation of transistor operation in subthreshold.

4.1 Single Transistor Circuits

MOSFETs are cool. They really have multiple modes of function with their different "regimes", which allow them to perform several functions. One thing to keep in mind before starting to look at circuits is that, as a rule of thumb, circuits are always drawn in a way to have the higher potential on top and the lowest potential on the bottom of the circuit. So you'll pretty much always see ground at the bottom and V_{dd} at the top, hopefully this will help you read through the circuits more easily.

4.1.1 The Current Source

The simplest function of a MOSFET, it is obtained by holding source, drain and gate voltage at constant values. As long as the difference between the drain and source voltages is larger than approximately $4U_T$, and that $V_{gs} < 0.7V$ the nFET in **saturation and subthreshold** reduces to:

$$I_{ds} = I_{n0} e^{\frac{\kappa_n V_g - V_s}{U_T}} \quad (57)$$

Remember from our previous discussion of second order effects (see section 3.9 that this is a First-Order approximation (i.e., it neglects Second-Order effects such as the Early Effect). This basically yields the assumption that the drain current is independent of the drain voltage in Saturation. Note that most often, current source encountered in circuits are pFETs rather than nFETs.

³⁹VLSI stands for Very-Large-Scale-Integration

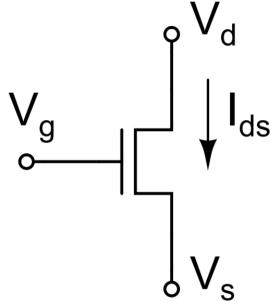


Figure 47: nFET Current Source. Adapted from Lecture Notes.

4.1.2 Linear Resistor

With some calculation and a lot of approximation, we can reach Ohm's relation, back from the known I_{ds} equation. This means we have a value that links between voltage and current: resistance. Let's start with the subthreshold equation, but without assuming saturation:

$$I_{ds} = I_{n0} e^{\frac{\kappa_n V_g}{U_T}} \left(e^{\frac{-V_s}{U_T}} - e^{\frac{-V_d}{U_T}} \right) \quad (58)$$

With some rearranging, we get:

$$I_{ds} = I_{n0} e^{\left(\frac{\kappa_n V_g}{U_T} - \frac{V_d + V_s}{2U_T}\right)} \left(e^{\frac{V_d - V_s}{2U_T}} - e^{-\frac{(V_d - V_s)}{2U_T}} \right) \quad (59)$$

Lucky us, the second term looks like a sinh! (As $\sinh(x) = \frac{e^x - e^{-x}}{2}$). We thus reach:

$$I_{ds} = 2I_{n0} e^{\left(\frac{\kappa_n V_g}{U_T} - \frac{V_d + V_s}{2U_T}\right)} \sinh\left(\frac{V_d - V_s}{2U_T}\right) \quad (60)$$

And here comes the approximation that will please mathematicians: If $V_d - V_s$ is small enough (i.e., if we are in the ohmic region), we can approximate the sinh with Taylor Series while neglecting higher order effects:

$$I_{ds} \approx I_{n0} e^{\left(\frac{\kappa_n V_g}{U_T} - \frac{V_d + V_s}{2U_T}\right)} \frac{V_d - V_s}{U_T} \quad (61)$$

Now if we freeze V_d , V_s and V_g , our MOSFET suddenly starts acting like a linear resistor with the following relation:

$$R = \frac{U_T}{I_{n0}} e^{\frac{V_d + V_s}{2U_T} - \kappa_n \frac{V_g}{U_T}} \quad (62)$$

I think it's really sad that I took the time to write all these equations and we don't even need to know them by heart. I hope my suffering got you to at least understand the gist of what using a transistor as a resistor is about. Though there remains an important question:

Why should we bother implementing a resistor with a Transistor instead of just using an actual resistor?

Good question, and I'm glad you asked. Well in standard CMOS circuits, you're dealing with very very small material, and resistors of this size are very unpractical. It also makes it a lot easier for building purposes, as you don't need to add another type of components to the circuit - Transistor can do everything you need! Another important advantage is that we can change the value of the resistance of the transistor, whereas it is fixed with a physical resistor. Though in most cases, we use a Transconductance amplifier to implement a resistance, but more on this in a dedicated chapter!

4.1.3 Non Linear Current-Voltage / Voltage-Current Converter

Sometimes you really need a Gin-Tonic, but all you have at home for you and your friends is juice. Conversely, sometimes you really need a nice juice to forget your difficult night out of the day before, but all you have is Gin left-overs. It would be so convenient if you could easily turn one into the other? It's the same in electrical circuits, where you sometimes need voltage for a given operation, but all you have to give is current, or the opposite. Well, with Transistors, you can convert a current into a voltage, and vice versa! Remember from ?? that MOSFET operating in saturation and subthreshold can generate a drain current which is an *exponential* function of V_{gs} . Now if you take a current as the input signal, we can isolate V_g or V_s and make it the output signal. This was a bit counter intuitive to me at first: we've always looked at voltages as the first thing to apply to obtain current. But it does make sense that if you *force* a current into a transistor, the voltage will have to follow in order to satisfy their expected behaviour. In subthreshold, we can re arrange ?? and isolate V_s or V_g as follows:

$$V_s = \kappa_n V_g - U_T \log\left(\frac{I}{I_{n0}}\right), \quad (63)$$

$$V_g = \kappa_n^{-1}(V_s + U_T \log\left(\frac{I}{I_{n0}}\right)) \quad (64)$$

4.1.4 Diode Connected Transistors

This one can be tricky at first, and it's very important to understand it properly.

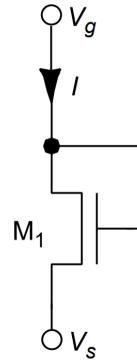


Figure 48: Diode Connected nFET. Adapted from Lecture Notes.

In this circuit, we connect the gate to the drain: $V_g = V_d$. This reduces the transistor to a two terminal device, similar to a diode, hence its name diode connected. Understanding the property of the circuit is actually simpler with the help of the equations than words. If we go from our full subthreshold equation 24:

$$I_{ds} = I_{n0} e^{\frac{\kappa_n V_g}{U_T}} (e^{\frac{-V_s}{U_T}} - e^{\frac{-V_d}{U_T}}) \quad (65)$$

we can rearrange as follows by replacing V_d with V_g :

$$I_{ds} = I_{n0} e^{\frac{\kappa_n V_g}{U_T}} (e^{\frac{-V_s}{U_T}} - e^{\frac{-V_g}{U_T}}) = I_{n0} (e^{\frac{\kappa_n V_g - V_s}{U_T}} - e^{\frac{\kappa_n V_g - V_g}{U_T}}) \quad (66)$$

$$I_{ds} = I_{n0} e^{\frac{\kappa_n V_g - V_s}{U_T}} - e^{\frac{\kappa_n V_g - V_g}{U_T}} \quad (67)$$

Remember, we love assumptions and simplifications of calculations here. So we keep on assuming that $\kappa_n \approx 1$

$$I_{ds} \approx I_{n0} e^{\frac{\kappa_n V_g - V_s}{U_T}} - e^{\frac{V_g - V_g}{U_T}} \quad (68)$$

and thus reach back the familiar saturation equation:

$$I_{ds} \approx I_{n0} e^{\frac{\kappa_n V_g - V_s}{U_T}} \quad (69)$$

So now, when a transistor is diode connected, it actually *necessarily operates in saturation*, it's not an assumption anymore (well it always kinda is, but less than before :)). Now the critical thing to understand is that the first thing happening here is current flowing into a transistor (just imagine a current source as described before). The current flowing creates a feedback loop between the drain and the charge where both automatically adapt to match each other and work saturation. So the current is what sets the gate voltage! This is very important as it is a clever way to adjust gate voltage from the current, which is typically not possible as there is infinite impedance between the channel (source, drain and well) and the gate.

4.2 Two Transistor Circuits

Before you go onto this section, make sure you fully grasp the idea behind the single transistor circuits.

4.2.1 Current Mirror

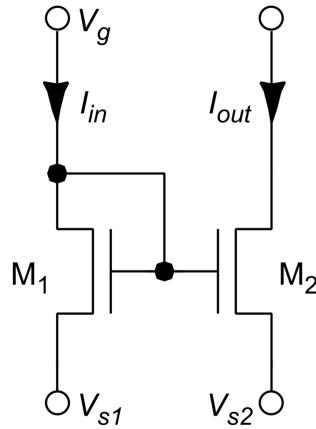


Figure 49: Current Mirror. Adapted from Lecture Notes.

In this circuit, the output current is a *mirrored copy* of the input current. What does this mean? Suppose both nFETs M1 and M2 1) have the same source voltage, 2) have the same size (so weird difference due to device mismatch don't apply) and 3) are in saturation. Remember from the previous section that the current I_{in} flowing through the Diode-Connected transistor M1 sets the gate voltage. You can also see that both M1 and M2 share the same gate voltage. Thus, given the assumptions taken, we reach a circuit where I_{in} sets I_{out} of M2. What's most interesting is that we can also scale the output current I_{out} by setting difference source voltages V_{s1} and V_{s2} , or by having different transistor sizes. Let's derive the equation when we have different source voltages:

$$\text{For M1 : } I_{in} = I_{n0} e^{\frac{\kappa_n V_g - V_{s1}}{U_T}} = I_{n0} e^{\frac{\kappa_n V_g}{U_T}} e^{\frac{-V_{s1}}{U_T}} \quad (70)$$

$$\text{For M2 : } I_{out} = I_{n0} e^{\frac{\kappa_n V_g - V_{s2}}{U_T}} = I_{n0} e^{\frac{\kappa_n V_g}{U_T}} e^{\frac{-V_{s2}}{U_T}} \quad (71)$$

$$I_{n0} e^{\frac{\kappa_n V_g}{U_T}} = \frac{I_{in}}{e^{\frac{-V_{s1}}{U_T}}} = \frac{I_{out}}{e^{\frac{-V_{s2}}{U_T}}} \quad (72)$$

$$\text{Thus, } I_{out} = I_{in} e^{\frac{V_{s1} - V_{s2}}{U_T}}, \text{ with Gain } M = e^{\frac{V_{s1} - V_{s2}}{U_T}} \quad (73)$$

We can also write the equation for same source voltage but different transistor size, without going into the derivation we reach:

$$I_{out} = MI_{in} \text{ with Gain } M = \frac{W_2/L_2}{W_1/L_1} \quad (74)$$

P-Type Current Mirror

A common exam question is to draw the P-type equivalent circuits of circuits we've learned. It also turns out that the P-type current mirror will be particularly important when studying the transconductance amplifier, so let's take some time to briefly study its behaviour.

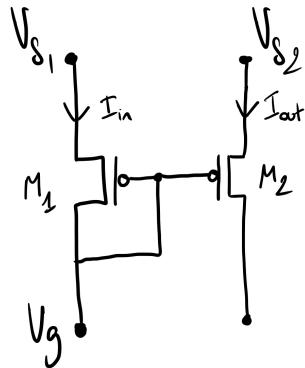


Figure 50: P-Type Current Mirror. Adapted from my brain lol.

The P-type current mirror works just like the N-type, but in reverse. In the N-type current mirror, we *diode connect* the drain of the first transistor, *at higher potential*, to the gate, thereby ensuring saturation. In an PFET, this is the opposite, we *diode connect* the drain of the first transistor, *at lower potential*, to the gate. This tiny difference ends up yielding the exact same output equation for I_{out} .

4.2.2 Intrinsic Voltage Gain

In the previously derived current mirror, a gain has been evidenced between the input current and the output current: that means that we manage to scale the *current* by a certain constant value that we have control over when designing our circuit. This is very useful in a lot of different applications and circuits that will be derived next. It also is very useful to be able to scale *voltage*!

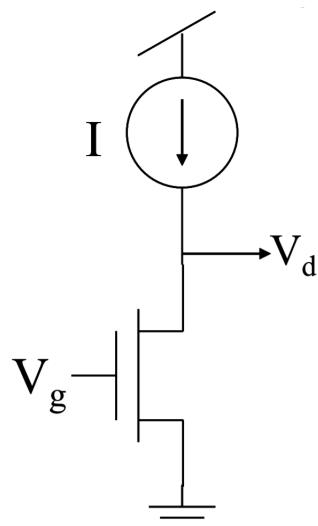


Figure 51: Intrinsic Voltage Gain with Transistors. Adapted from Lecture Notes.

The first thing to note in this figure is the current source. We've seen them before, but this one is slightly different: it's a pFET current source (remember I made a comment about pFET current source being most often used!). The source voltage V_s is at V_{dd} and the drain voltage is V_d . Current flows from positive to negative, so from top to bottom (current, not electrons - thanks again Benjamin Franklin). Note again that this allows to have a *constant* current flowing. We can evaluate the gain of this circuit as follows:

$$\text{Gain } A = \frac{\partial V_d}{\partial V_g} = \frac{\partial I}{\partial V_g} \frac{\partial V_d}{\partial I} = \frac{g_m}{g_d} {}^{40} = \frac{\kappa V_E}{U_T} \quad (75)$$

The voltage gain is thus set by tweaking κ , Early voltage V_E and thermal voltage U_T .

4.2.3 Source Follower

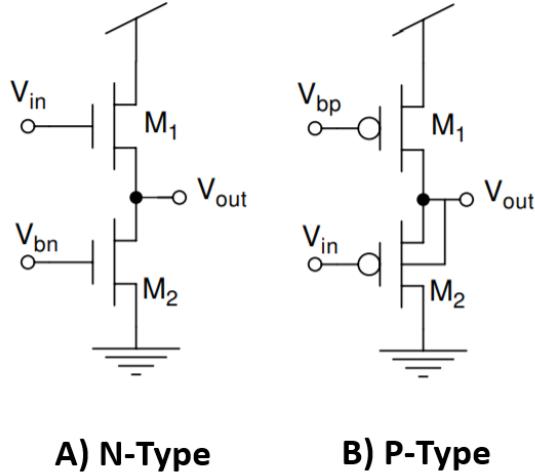


Figure 52: Source Follower circuits. A) N-Type Source Follower. B) P-Type source Follower. Adapted from Lecture Notes.

The source-follower circuit linearly transforms a voltage at a high-impedance input terminal into a voltage at a lower-impedance output terminal, such that the output signal is able to drive larger loads than the input signal. It is constructed by connecting a fixed current source to the source of a MOSFET operated in saturation.

N-Type Source Follower:

In figure 13.A, we can see a N-Type source follower. M_2 is the NFET current source and M_1 is the input transistor. The input voltage V_{in} is applied to the gate voltage of M_1 and the output voltage V_{out} is the source voltage of M_1 . In the subthreshold domain, assuming both transistors are in saturation regime, we can derive the following equations describing the behaviour of the circuit:

$$I_{M1} = I_{n0} e^{\kappa_n V_{in}/U_T - V_{out}/U_T} \quad (76)$$

$$I_{M2} = I_{n0} e^{\kappa_n V_{bn}/U_T - V_{s_{M2}}/U_T} = I_{n0} e^{\kappa_n V_{bn}/U_T} \text{ as } V_{s_{M2}} = 0 \quad (77)$$

Because the two transistors are connected in series, $I_{M1} = I_{M2}$, we can therefore reach by rearranging:

$$V_{out} = \kappa_n (V_{in} - V_{bn}) \text{ with } V_{out} > 4U_T \text{ to keep } M_2 \text{ in saturation.} \quad (78)$$

What we get is thus an output voltage which follows the input voltage, and can be scaled with the gate voltage from the current source that we connect it with.

⁴⁰We have defined transistor conductance in chapter 3.

P-Type Source Follower:

In figure 13.B, we can see a P-Type source follower. The logic and objective of the circuit are pretty much the same as the NFET version, but with a PFETs. M_1 is the PFET current source and M_2 is the input transistor. An important difference here is that V_{out} is connected to the bulk of M_2 . Remember from the 3.7 section on bulks, wells and biasing that MOSFETs are typically biased to V_{dd} or V_{ss} at their bulk. Here, they are biased to V_{out} . It is possible to get around the κ reduction factor in the transfer characteristic, if the bulk potential of the input MOSFET can be controlled independently. As mentioned previously, in a CMOS process this independence is possible for only one type of MOSFET: The one that sits in a well with opposite doping from the substrate.

$$I_{M1} = I_{p0} e^{-\kappa_p(V_{bp}-V_{dd})/U_T} \quad (79)$$

$$I_{M2} = I_{n0} e^{-\kappa_p(V_{in}-V_{out})/U_T} \quad (80)$$

Because the two transistors are connected in series, $I_{M1} = I_{M2}$, we can therefore reach by rearranging:

$$V_{out} = (V_{dd} - V_{bp}) + V_{in} \text{ with } V_{out} < V_{dd} - 4U_T \text{ to keep } M_1 \text{ in saturation.} \quad (81)$$

Notice how we got rid of the κ here, where V_{out} is now linear with the input. This variation of the source follower is called a *unity gain* source follower. What's the point of this circuit though? Why would you want to have an output equal to the input? Why should you go through the trouble of building identity. The reason, according to Giacomo, is to **decouple** the left side from the right side. You might have noise, load or capacitance on one side and you don't want to influence the rest of the circuit with that and keep it clean. You could also use it, as in the previous nFET source follower, use it to shift by some value the V_{out} compared to V_{in} .

4.3 Three (and more) Transistor Circuits

In the following, we will turn to some slightly more complex circuits in order to introduce the principles of the transconductance amplifier, which is used in a variety of circuit configurations in analog circuit design.

4.3.1 The differential pair

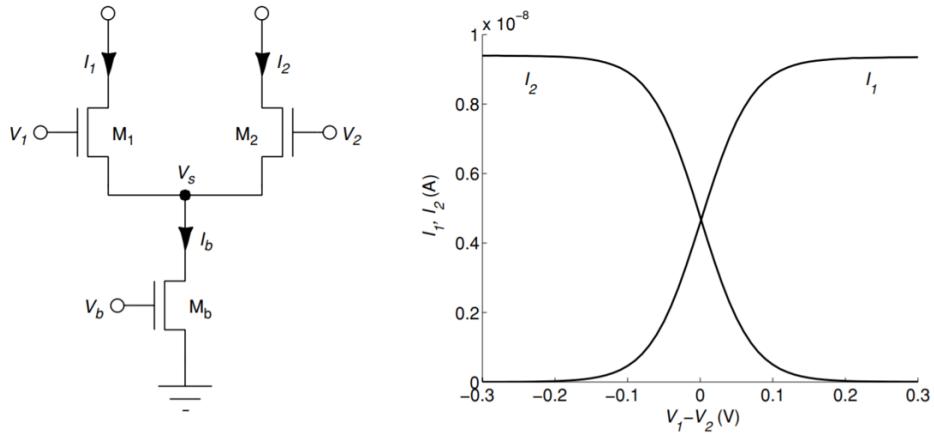


Figure 53: (A) Differential Pair Circuit. (B) Differential Pair output currents on differential input voltage. Adapted from textbook.

The differential pair has the same basic structure as the source follower, except that the current source, also called the *bias current* I_b is now shared by two MOSFETs M_1 and M_2 whose sources are connected to the drain of the bias MOSFET M_b . The sharing of the current between M_1 and

M_2 depends on their respective gate voltages V_1 and V_2 . If all MOSFETs are operated below threshold and in saturation and we assume that M_1 and M_2 have the same subthreshold slope factor κ_n , we obtain the following equations (assuming all transistors work in saturation):

$$I_1 = I_0 e^{\frac{\kappa V_1 - V_S}{U_T}} \quad (82)$$

$$I_2 = I_0 e^{\frac{\kappa V_2 - V_S}{U_T}} \quad (83)$$

Because of Kirchoff's Current Law, $I_B = I_1 + I_2$, and $I_B = I_1 + I_2 = I_0 e^{\frac{\kappa V_b}{U_T}}$. Now we can rewrite I_b

$$I_b = I_0 e^{\frac{-V_S}{U_T}} (e^{\frac{\kappa V_1}{U_T}} + e^{\frac{\kappa V_2}{U_T}}) \quad (84)$$

With some algebra (by factoring out $e^{\frac{-V_S}{U_T}}$), we can reach the elegant rewriting of I_1 and I_2 as a function of the two input voltages:

$$I_1 = I_b \frac{e^{\kappa_n V_1 / U_T}}{e^{\kappa_n V_1 / U_T} + e^{\kappa_n V_2 / U_T}} \quad (85)$$

$$I_2 = I_b \frac{e^{\kappa_n V_2 / U_T}}{e^{\kappa_n V_1 / U_T} + e^{\kappa_n V_2 / U_T}} \quad (86)$$

Now if we take the difference between I_1 and I_2 :

$$I_1 - I_2 = I_b \frac{e^{\frac{\kappa V_1}{U_T}} - e^{\frac{\kappa V_2}{U_T}}}{e^{\frac{\kappa V_1}{U_T}} + e^{\frac{\kappa V_2}{U_T}}} = I_b \tanh\left(\frac{\kappa(V_1 - V_2)}{2U_T}\right) \quad (87)$$

The dependence of these two currents on the difference of the two input voltages is shown in Figure 53.B. The curves have a sigmoidal shape ⁴¹. Actually, if you re-arrange any of these two equations, you obtain exactly the form of a sigmoid function ⁴². They are almost linear for small voltage differences and saturate at I_b for large voltage differences. We saw that the difference between I_1 and I_2 gives a tanh function. Such compressive nonlinearities (the sigmoid and tanh) are very useful for the implementation of different functions, especially in the context of neural networks. What makes the circuit even more useful is the fact that to a first approximation (neglecting the Early effect), the output currents depend only on the difference of the input voltages: The circuit has a small *common-mode* sensitivity. Here is a summary of things about Differential pairs that you want to keep in mind if you get to talk about it in the exam:

- These equations work assuming all transistors operate in subthreshold and saturation.
- The assumption that M_b works in saturation should normally not be taken. We should use the full equation to evaluate the current I_b . However, in this chapter we chose to just assume this. We will, in the next chapter on Transconductance amplifier, specifically derive the condition that need to be satisfied for M_3 to operate in saturation (that is $V_S > 4U_T$, which will yield a specific V_1 , V_2 and V_b relation to satisfy). Don't worry about this too much for now, it will make sense in the next chapter.
- The dependence of these two currents on the difference of the two input voltage gives nice and smooth sigmoid functions, with a linear I-V relationship for small voltage differences. You can use this property to implement programmable resistors in the circuit!
- Voltages are differential rather than absolute quantities - this is very useful, for things such as cancelling out noise.
- If you take the difference between I_1 and I_2 , you reach a hyperbolic tangent function. This also gives a linear I-V relationship for small voltage differences, which is a property we will use in the transconductance amplifier.

⁴¹Sigmoid are very important functions in Neural Networks. You can read a bit more about them here: <https://machinelearningmastery.com/a-gentle-introduction-to-sigmoid-function/>

⁴² $I_1 = I_b \frac{e^{\kappa_n V_1 / U_T}}{e^{\kappa_n V_1 / U_T} + e^{\kappa_n V_2 / U_T}} = I_b \frac{1}{1 + e^{\frac{\kappa(V_2 - V_1)}{U_T}}}$, which is a sigmoid function with $x = -\frac{\kappa(V_2 - V_1)}{U_T}$

4.3.2 The current correlator

The current correlator measures the correlation between unidirectional input currents, whereas the bump circuit (which we'll look at just after that) measures the similarity or dissimilarity of input voltages. Both of these circuits have been used in many analog VLSI designs. For example, these circuits have been used in a stereoscopic vision system (Mahowald, 1994) to disambiguate between real and false targets. Both circuits are extensively discussed in Tobi's 1993 paper⁴³

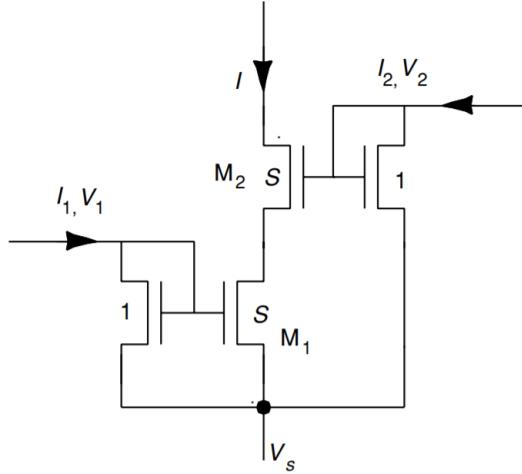


Figure 54: Current Correlator. Adapted from textbook.

Carver Mead recognized that in subthreshold operation, the current-correlator circuit in Figure 54 computes a measure of the correlation between its two input current I_1 and I_2 . Intuitively, the series-connected transistors perform an analog logic **AND** like computation. If either of the gate voltages on these series connected transistors is low, then the output current is shut off. Conversely, if both of the input voltages are high, then the output current is large. In the intermediate regions, the circuit computes an approximation to the product of the input currents. Some important details: M_1 is in the Ohmic region, and M_2 is in saturation. We can thus derive the equations of function for this circuit: As M_1 and M_2 are connected in series, the same current I_{out} flows through them.

$$\text{For } M_1 : I_{out} = I_0 e^{\kappa V_1 / U_T} (e^{-V_S / U_T} - e^{-V / U_T}) = I_0 e^{\kappa V_1} (1 - e^{-V}) \text{ as } V_S = 0 \quad (88)$$

$$\text{For } M_2 : I_{out} = I_0 e^{\kappa \frac{V_2 - V}{U_T}} \quad (89)$$

Here V is the drain voltage of M_2 and the source voltage of M_1 . For simplicity, let's ignore U_T in the following equations which cancels out anyways. We can factor the last equation as follows:

$$e^V = \frac{I_0 e^{\kappa V_2}}{I_{out}} \quad (90)$$

From simple saturation equation, we can derive I_1 and I_2 :

$$I_1 = I_0 e^{\kappa V_1}; \quad I_2 = I_0 e^{\kappa V_2} \quad (91)$$

After some very tedious algebra and rearranging, we find back the elegant equation:

$$I_{out} = \frac{I_1 I_2}{I_1 + I_2} \quad (92)$$

⁴³<https://www.ini.uzh.ch/tobi/anaprose/bump/index.php>

4.3.3 The Bump-antibump circuit

The bump anti bump circuit is a circuit invented by Tobi⁴⁴, that also aims at computing the similarity and difference between two input signals. While we do not need to know the details of the circuit, it is important (and interesting) to describe qualitatively what computation this rather complex circuit does.

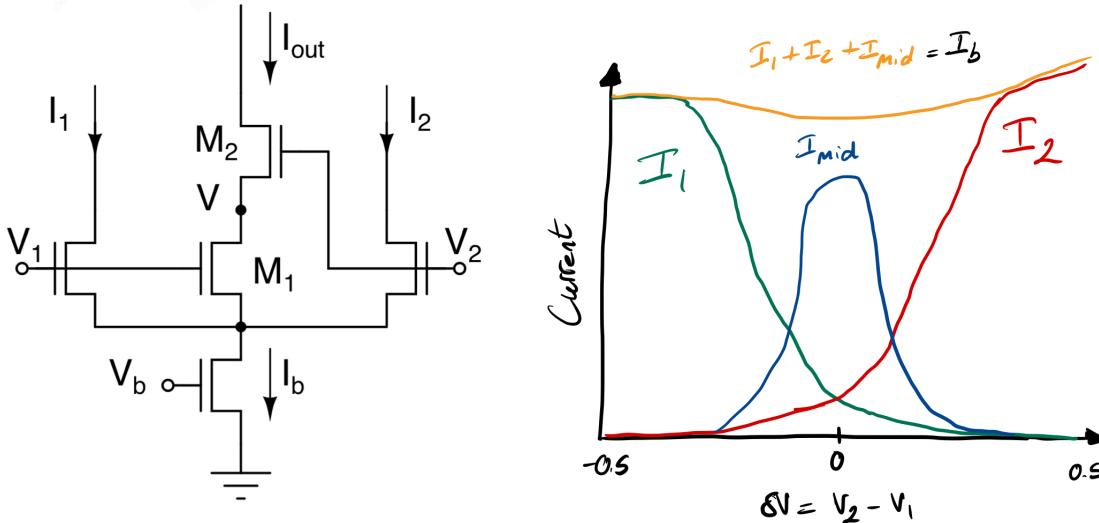


Figure 55: (a) Bump-Antibump circuit. (b) Output characteristics of bump-antibump. circuit Adapted from Lecture Notes. Output characteristics is Yassine's drawing, inspired from Hector the Saviour's explanation. Thank you Hector.

Figure 55.A shows the bump-anti bump circuit, which looks a lot like the current correlator when you pay attention, and it is not so dissimilar on a functional level. It has 3 outputs: I_1 , I_2 and I_{out} , also often called I_{mid} . Output I_{out} is the *bump* (as can be seen from the graph), and I_1 and I_2 form, combined, the *anti-bump* of the circuit. We can intuitively reason through the circuit: The three current must sum to the bias current I_b due to Kirchoff's Current Law. Hence, voltage V_C follows the higher of V_1 or V_2 . The series connected transistors M_1 and M_2 form the core of the same analog current correlator that is used in the current-correlator. When $\delta V = 0$, current flows through all three legs of the circuit. When δV increases, the common-node voltage V_C begins to follow the higher of V_1 or V_2 , thereby shutting off I_{mid} . This is because the transistor whose gate is connecter to the lower of V_1 or V_2 shuts off! Indeed, if V_1 increases, I_1 increases, but it can never increase beyond I_b as $I_b = I_1 + I_2 + I_{mid}$. So we reach a point where I_1 gets all the current with a high V_1 , not letting anything through I_2 . This leads V_2 to adapt itself in order to virtual shut off I_2 , and I_{mid} with it as it's the gate voltage of M_2 . Same applies in reverse when increasing V_2 and I_2 . So I_{mid} is max when $\delta V = 0$ and shut off with a significant δV . Conversely, I_1 and I_2 respectively follow V_1 and V_2 .

We could write the equations that describe how I_1 , I_2 and I_{mid} behave as a function of input voltages, but it's just fancy equation that you won't remember and won't be asked about in the exam. Just remember the intuition behind the circuit!

4.4 Laboratory : Static Circuits

In this fourth lab , we are going to use our knowledge of subthreshold transistors to characterize two circuits : the differential pair and the Bumb antibumb.

To measure small current (in our case from 1 pA to 10 nA) we are going to use a converter circuit, which will map current to frequency (C2F). Be aware of this because later on we need to map back our results to current.

⁴⁴<https://www.ini.uzh.ch/tobi/anaprose/bump/index.php>

Today, you will also be first exposed to a Multiplexer and demultiplexer (mux/demux) circuit. The transistors we use are very tiny and numerous, but the input-output pins and C2F converters are unfortunately spacious. Therefore, we need mux/demux to select the circuits that we want at the beginning of each lab.

Another functional circuit we use is the Bias Generator. It will make sure that all transistors work in the subthreshold regime by mirroring a current to the circuits that we need.

Thus, we have three functional circuits that can let us advance to more complicated experiments

- . These are : - C2F converter - Multiplexer and demultiplexer - Bias Generator

They are here to help you, just be aware of them.

4.4.1 N-FET differential pair circuit (NDP)

First, we need to calibrate the C2F converter and the bias generator. To calibrate the output of C2F, we set a high voltage differential between V_1 and V_2 (i.e. around 0.4V) and we loop through a range of bias currents. We then read the output (which are frequencies) on the I_1 and I_2 . This way we have a dictionary to refers to when we map frequency back to current in a later stage.

The interesting part starts now. We fix a common voltage between V_1 and V_2 and we loop through a range of V_1 and V_2 that respect this common voltage constraint.

I_1 becomes dominant when the differential $V_1 - V_2$ is higher than -0.1V. The two output current are symmetrical, whereas the difference between the two is a sinusoidal. Overall, the current in the circuit is always equal to the bias current I_b .

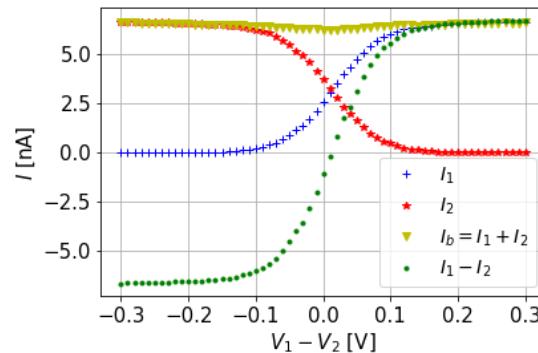


Figure 56: Interpolated differential pair currents plotted over the voltage difference $V_1 - V_2$

If we increase the bias or the common mode, we can increase the current flowing in the circuit, hence, also the maximum ampere of I_1 and I_2 .

In the range of linearity, we observe that the offset voltage is slightly greater than zero because the two transistors are not exactly the same (as it is in theory assumed by the model). We note that the scale of the linearity range is the thermal voltage because the transistor is running under the weak inversion regime. We speculated that if the regime was strong inversion, the overdrive would have determined this range, but it turned out we were wrong.

4.4.2 Bumb antibump

The process to get results in the Bumb Anti Bumb circuit is comparable to the one used before . We calibrate the channels and then we extract results varying V_1 and V_2 under a common voltage constraint (i.e. 0.9V).

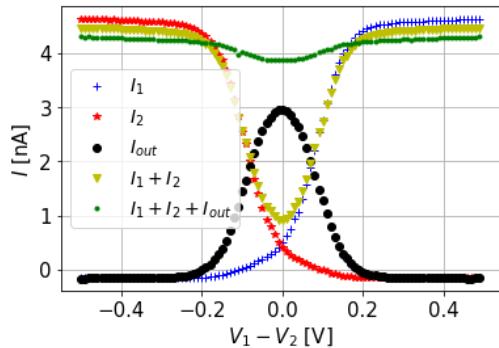


Figure 57: Interpolated Bumb AntiBumb plotted over the voltage difference $V_1 - V_2$

What is important to notice, is that the sigma of the Guassian distribution I_{out} (sorry for the approximation, but I see Bell curves everywhere) it's proportional to the current I_b . Keep this in mind, but don't bring it up near to Tobi because it awakens bad memories to him ;).

The behavior of I_{out} is very well defined because we are interpolating frequencies outputs. Therefore, we can easily fit a quadratic function to the sum $I_1 + I_2$ and a linear function for the individual components.

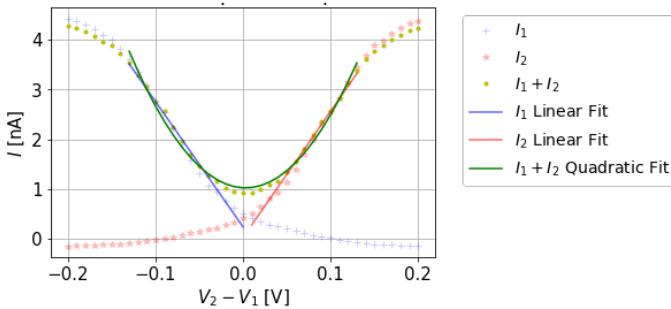


Figure 58: Quadratically fitted interpolated Bumb AntiBumb plotted over the voltage difference $V_1 - V_2$

4.5 Test yourself

You should be able to answer the following things:

- Explain what the Early effect is, why it's important and when one should pay attention to it.
- Understand how a diode-connected transistor function, both through nFET and pFET!
- Draw a current mirror, and qualitatively explain how it works and the assumptions taken to derive the behaviour.
- Draw a differential pair, and qualitatively describe its functioning. sketch the I-V relationship between the input voltage difference and output currents. Derive its equation.
- Draw a current correlator, and qualitatively describe its functioning. Know its equation.
- Qualitatively describe the functioning of a bump antibump circuit

5 The Transconductance Amplifier

If a whole lecture is dedicated to this specific circuit, it is for a good reason. It is a building block of many different electrical circuits, within and beyond Neuromorphic Engineering. You may have heard of its cousin the Operational Amplifier (commonly called OpAmp), which is not that different. In this chapter, we'll first look at the architecture of this circuit, and then look at precise function with different case scenario. We'll finish by briefly looking at another circuit that is built from the transconductance amplifier: the Wide-output-range differential transconductance amplifier.

Here are things you should be comfortable with before starting to read through this chapter:

- Architecture and behaviour of the diode connected transistor (specifically P-Type)
- Architecture and behaviour of current mirror (specifically P-Type)
- Architecture and behaviour of differential-pair.
- Early Voltage and output conductance of transistors.

5.1 Architecture

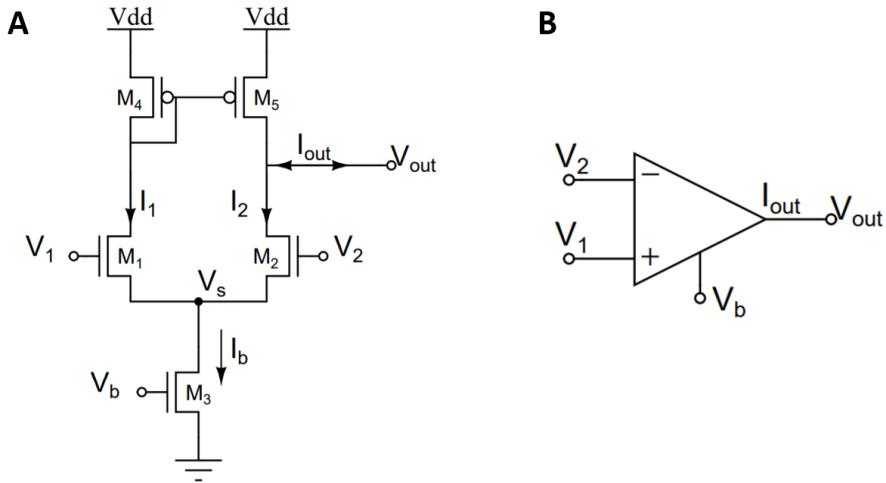


Figure 59: The Transconductance Amplifier. A) Transconductance Amplifier Circuit. B) Electrical symbol of the Transconductance Amplifier. Adapted from Lecture Notes.

In the Transconductance Amplifier circuit, shown in Figure 59.A, the first thing to pay attention to is that it is a combination to two circuits we have studied before: the differential pair (on the lower end) and a P-Type current mirror (on the higher end). Let's remind ourselves of the important equations and assumptions of the diff pair and the P-type current mirror.

The differential pair Remember that the differential pair implements a difference between the two input voltages V₁ and V₂ which translates to the two output currents I₁ and I₂. Here are the equations:

$$I_1 = I_b \frac{e^{\kappa_n V_1 / U_T}}{e^{\kappa_n V_1 / U_T} + e^{\kappa_n V_2 / U_T}} \quad (93)$$

$$I_2 = I_b \frac{e^{\kappa_n V_2 / U_T}}{e^{\kappa_n V_1 / U_T} + e^{\kappa_n V_2 / U_T}} \quad (94)$$

We previously assumed that all transistors work in subthreshold and saturation. Remember also the sigmoid behaviour of the two output currents, as well as the hyperbolic tangent (tanh) behaviour of the difference between the two currents.

The P-type current mirror We have shown in the previous chapter that the P-Type current mirror behaves as follows:

$$I_{out} = I_{in} e^{\frac{V_{s1}-V_{s2}}{U_T}} \quad (95)$$

In our case, both V_{S_1} and V_{S_2} are V_{dd} . We therefore have a unity gain: $I_{out} = I_{in}$.

5.2 Transconductance Amplifier Function

We can now start to analyze the dynamics when both the current mirror and the diff pair are connected to form a transconductance amplifier.

5.2.1 Let's assume everything is in Saturation

Let's first start with, as always, some assumptions and consequent observations.

- M_1 works in saturation: $I_1 = I_0 e^{\frac{\kappa V_1 - V_S}{U_T}}$
- M_2 works in saturation: $I_2 = I_0 e^{\frac{\kappa V_2 - V_S}{U_T}}$
- M_3 works in saturation: $I_b = I_0 e^{\frac{\kappa V_b}{U_T}}$
- M_4 and M_5 are also in saturation. We can therefore apply the P-Type current mirror equation reached previously.

Now let's see what we can reach with this. We noted previously that our current mirror had a unity gain, this means that I_1 flows through M_5 (since it is mirroring the current of M_4). Using Kirchoff's current law, we can now establish that: $I_{out} = I_1 - I_2$. If we take the expression derived in the assumptions for I_1 and I_2 , we reach with some tedious algebra:

$$I_{out} = I_1 - I_2 = I_b \frac{e^{\frac{\kappa V_1}{U_T}} - e^{\frac{\kappa V_2}{U_T}}}{e^{\frac{\kappa V_1}{U_T}} + e^{\frac{\kappa V_2}{U_T}}} = I_b \tanh\left(\frac{\kappa(V_1 - V_2)}{2U_T}\right) \quad (96)$$

This yields the following dynamic:

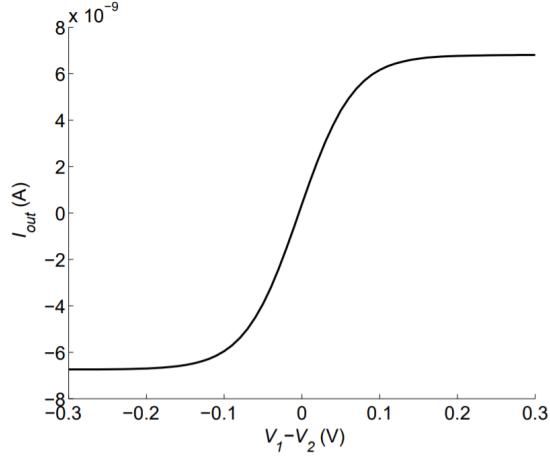


Figure 60: The Transconductance Amplifier ouptput current as a function of the difference between V_1 and V_2 . This is assuming that all transistors, especially M_3 are in saturation. Adapted from Lecture Notes.

Importantly, we note that for small differential voltages ($|V_1 - V_2| < 200\text{mv}$), the tanh relationship is approximately linear and our equation can be reduced to:

$$I_{out} \approx g_m(V_1 - V_2) \text{ with } g_m = \frac{I_b \kappa}{2U_T} \quad (97)$$

The term g_m is the **transconductance** of the amplifier. It quantifies how much the potential difference will influence the output current: $g_m = \frac{\partial V_{in}}{\partial I_{out}}$. It has the dimensions of a conductance (1/Ohms). To increase g_m , you can only increase I_b . Usually, a high transconductance is more desirable, as it leads to a higher out current for the same current. The issue is that to increase I_b , you need a higher V_b , which means consuming more energy. As Shih-Chii says: there is no free lunch! Because the output current is measured at a terminal which is different from the pair across which the input voltage difference is applied, we can also define another term (which we've defined before!): the **output conductance** $g_d = \frac{\partial I_{out}}{\partial V_{out}}$. Instead of measuring the change in current as a function of *input* voltage as with g_m , we measure the change in current as a function of the change in *output* voltage. We reach the familiar expression:

$$g_d = -\frac{\partial I_{out}}{\partial V_{out}} \approx \frac{I_b}{V_E} \quad (98)$$

where V_E is the Early Voltage of M_2 and M_5 , assumed equal.

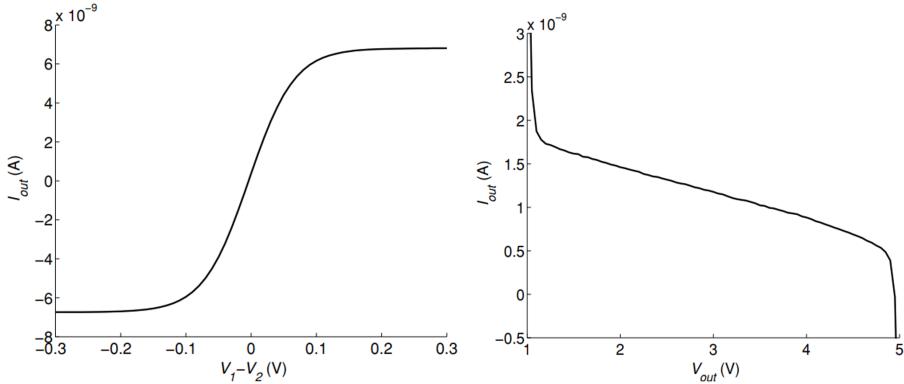


Figure 61: A) Transconductance of Transamp g_m . B) Output conductance of Transamp g_d .

5.2.2 Let's stop assuming that everything is in Saturation

Now, let's consider the case where we cannot assume that M_3 operates in saturation (does not mean it is not in saturation, just that we cannot assume it). This yields a different behaviour. We assume that the rest is still in saturation. We therefore have to write the full equation for current I_b .

$$I_b = I_0 e^{\frac{\kappa V_b}{U_T}} \left(1 - e^{\frac{-V_s}{U_T}} \right) \quad (99)$$

From the differential pair, we know that $I_b = I_1 + I_2$. Assuming that M_1 and M_2 are in saturation, we reach:

$$I_b = I_0 e^{\frac{\kappa V_b}{U_T}} \left(1 - e^{\frac{-V_s}{U_T}} \right) = I_0 e^{\frac{\kappa V_1 - V_s}{U_T}} + I_0 e^{\frac{\kappa V_2 - V_s}{U_T}} \quad (100)$$

Solving for e^{-V_s/U_T} , we reach with some, as always, tedious algebra:

$$e^{-V_s/U_T} = \frac{e^{\kappa V_b/U_T}}{e^{\kappa V_b/U_T} + e^{\kappa V_1/U_T} + e^{\kappa V_2/U_T}} \quad (101)$$

So we considered the case where we could not be certain of M_3 operating in saturation. We derived the subsequent equation, and now we want to find what are the conditions and implications of M_3 actually being in saturation. Confusing, I know. But just think of it as a more rigorous way to work with this circuit: we want M_3 to be in saturation but we can't assume it is, so let's figure out how to make it work the way we want. So: if we want M_3 to operate in saturation, we need

the V_{ds} of M_3 to be above $4U_T$, which means that $V_s > 4U_T$ and subsequently that $e^{-V_s/U_T} \ll 1$. Now let's do some maths:

$$e^{-V_s/U_T} = \frac{e^{\kappa V_b/U_T}}{e^{\kappa V_b/U_T} + e^{\kappa V_1/U_T} + e^{\kappa V_2/U_T}} \ll 1 \quad (102)$$

$$e^{\kappa V_b/U_T} \ll e^{\kappa V_b/U_T} + e^{\kappa V_1/U_T} + e^{\kappa V_2/U_T} \quad (103)$$

If we divide both sides by $e^{\kappa V_b/U_T}$, we reach:

$$1 + \frac{e^{\kappa V_1/U_T} + e^{\kappa V_2/U_T}}{e^{\kappa V_b/U_T}} \gg 1 \quad (104)$$

$$\frac{e^{\kappa V_1/U_T} + e^{\kappa V_2/U_T}}{e^{\kappa V_b/U_T}} \gg 1 \quad (105)$$

$$e^{\kappa V_1/U_T} + e^{\kappa V_2/U_T} \gg e^{\kappa V_b/U_T} \quad (106)$$

Remember that:

$$e^{-V_s/U_T} = \frac{e^{\kappa V_b/U_T}}{e^{\kappa V_b/U_T} + e^{\kappa V_1/U_T} + e^{\kappa V_2/U_T}} \quad (107)$$

Well from, that, we can derive V_s :

$$V_s = -\kappa V_b + U_T \ln(e^{\kappa V_b/U_T} + e^{\kappa V_1/U_T} + e^{\kappa V_2/U_T}) \quad (108)$$

Which, thanks to equation 106, yields:

$$V_s = -\kappa V_b + U_T \ln(e^{\kappa V_1/U_T} + e^{\kappa V_2/U_T}) \quad (109)$$

Now, taking the assumption that $|V_1 - V_2| > 4U_T$, we can finally turn into something useful:

$$V_s \approx \kappa(\max(V_1, V_2) - V_b) \quad (110)$$

Wait what? Where did the max come from? Yeah I asked myself the same question. It's no difficult to see: because the difference between $|V_1 - V_2| > 4U_T$, the expression inside the logarithm will converge to whichever of V_1 or V_2 is the biggest, hence the max function. Then it's just some factorizing to do. Well anyways, now we managed to derive what condition we need to satisfy to ensure that M_3 will be in saturation as it will have $V_S > 4U_T$:

$$\max(V_1, V_2) > V_b + \frac{4U_T}{\kappa} \quad (111)$$

Yes, all of this, just to know what kind of voltage we have to apply to V_1 , V_2 and V_B so that our transistor M_3 is in saturation. Now we can use the equation we derive in the previous section, without feeling bad about the fact we're just doing something unrealistic as we have too many assumptions. You see, assumptions aren't so bad in the end? And yes, I know, that was all very tedious, and probably unnecessary to derive here as we clearly won't have to do that for the exam. But don't get mad, I'm just trying to give you all the good info, and it's 2 in the morning so I am not thinking straight.

But what about the other Transistors? Before we move on, let's have a last look at the assumptions. So we just managed to derive the specific conditions to satisfy to ensure that M_3 operates in saturation. We, unfortunately, have to consider the saturation conditions for the other transistors as well. For practical purposes, M_1 and M_4 will always be in saturation because M_4 is *diode-connected* and the drain of M_1 is thus at a high voltage (THIS IS WORD TO WORD WHAT IS WRITTEN IN THE TEXTBOOK, BUT SHOULDN'T IT BE LOW VOLTAGE?). Now we need to look at M_2 and M_5 . So for these we're gonna have to work on it a bit.

- To keep M_5 in saturation, we need $V_{dd} - V_{out} > 4U_T$

- To keep M_2 in saturation, we need $V_{out} - V_s > 4U_T \equiv V_{out} > 4U_T + V_S$. We've just found V_S before, so let's use that, and we get

$$V_{out} > \kappa(\max(V_1, V_2) - V_b) + 4U_T \quad (112)$$

Huh... That's annoying. Our V_{out} also need to satisfy some condition which depends on V_1 , V_2 and V_b . This is called the min problem, because we manage to get in saturation only if the *minimum* V_{out} is greater than the condition above. And well, before we kinda also reached a min problem where we had saturation only if $|V_1 - V_2|$ was higher than $V_b + 4U_T/\kappa$. So you see that to get this circuit operated properly, you need to satisfy a lot of different things, which in practice are absolutely not trivial to satisfy.

- So how does this circuit work in practice? To be honest I am not sure myself and wouldn't mind having your input on it :).

5.2.3 Transconductance Amplifier as a Voltage Amplifier

This will be a brief comment on the subject. There is more to know than what we'll discuss, but we are not required to know the details. The transconductance amplifier can be used as a *differential* input voltage amplifier, where essentially: $V_{out} = A(V_1 - V_2)$. A being the gain of the circuit, relating the change in output compared to the change in input. Here is the transfer function, which uses some of our previously derived knowledge on transconductance and output conductance:

$$A = \frac{\partial V_{out}}{\partial(V_1 - V_2)} = \frac{\partial V_{out}}{\partial I_{out}} \frac{\partial I_{out}}{\partial(V_1 - V_2)} = \frac{g_m}{g_d} \approx \frac{\kappa V_E}{2U_T} \quad (113)$$

Here are key take aways from how this amplifier works:

- The open-circuit voltage gain A increases with Early Voltage and therefore with the length of the output transistors.
- Typical subthreshold values of A are between 100 and 1000.
- Because of the large voltage gain and transistor mismatch effects, the amplifier is usually used in a negative-feedback configuration. This mean that the circuit itself is too noisy to use its output as a function of its inputs, so it's either used as a comparator only (where the precise values don't matter) or by integrating a negative feedback that typically reduces the gain to create a more stable circuit.
- In open-voltage mode, it is used mainly as a comparator: V_{out} is "high" only when $V_1 > V_2$, and vice versa.

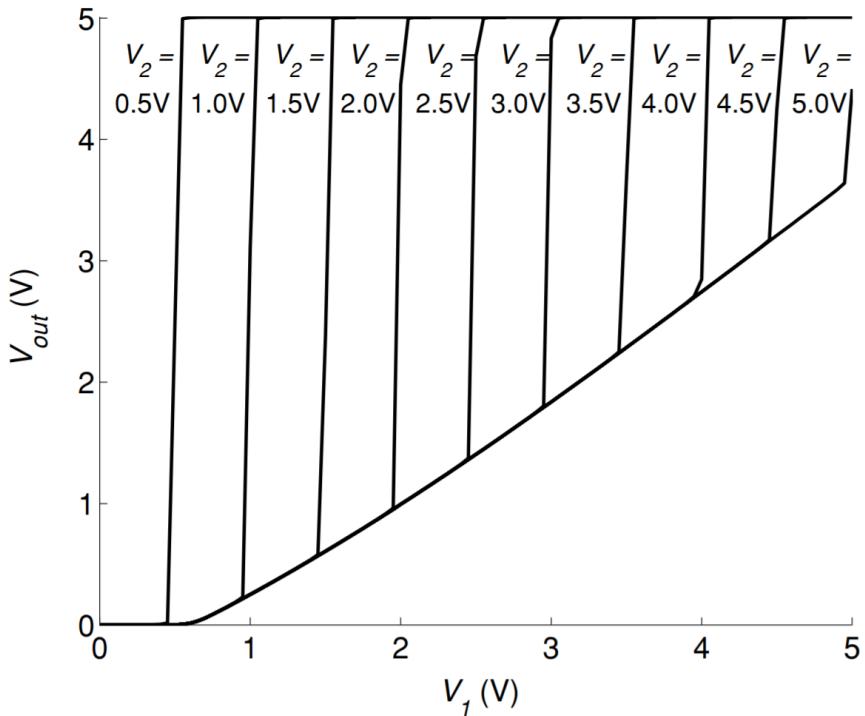


Figure 62: Voltage amplification characteristics of simple differential transconductance amplifier.
Adapted from Textbook.

Ok that was a lot for just one circuit. Let's summarize the key take aways:

- M_3 cannot be assumed to be in Saturation, so we used the full subtrehshold equation to find under what conditions M_3 could indeed be in saturation. We found that we needed to tune V_1 , V_2 and V_b to ensure that $\max(V_1, V_2) > V_b + \frac{4U_T}{\kappa}$ to be in saturation. If we do tune our voltages according to this equation, we are in saturation, and the I-V relationship derived in the previous section applies.
- Since V_{out} increases with increasing V_1 and decreases with increasing V_2 , the gate of M_1 is called the non-inverting input terminal and the gate of M_2 the inverting input terminal of the amplifier. It's the same with current, since $I_{out} = I_1 - I_2$, if V_1 is large, I_1 is large and V_{out} is large - hence it carrying the name of the non inverting terminal. Converse situation applies when V_2 is large.

6 Linear Systems Theory

This chapter will be some kind of exception to the rest of the chapters. Indeed, most of the Linear Systems' lecture covers basic of "Signals and Systems" - a field of its own, in order to properly introduce circuit response, steady state analysis and capacitive circuits which include time constants. When I attempted to write an exhaustive summary for this chapter, I realized four important things, which led me to organize the chapter the way I did:

1. It is simply impossible for anyone to understand the concepts introduced (delta function, convolution etc...) from only the introduction that was given in class. Signals and Systems is a paradigm of analysis on its own, and grasping the fundamentals properly takes, well, a lot of time. I took a whole course on this topic in my undergraduate degree and I still struggle to get the intuition behind convolution. Maybe I'm stupid.
2. There is no way for me to explain the basics of Signal and Systems without spending a whole month reviewing the topic on my own in order to find good analogies and ways of explaining it to the uninitiated.
3. Even if I did that, I'd never (in my wildest dreams) manage to do anything better than Alan V. Oppenheim in his classic textbook "Signals and Systems".⁴⁵
4. Last, and clearly not least, we are not required to know anything beyond Resistor-Capacitor circuit analysis (transfer functions etc..), which only require basic knowledge of how Laplace Transforms work.

I have thus decided to only cover the topics I mentioned in my last argument, that is: 1) a brief overview of exponential and Laplace Transforms and 2) full derivations of Resistor-Capacitor circuits transfer functions and dynamics. I have also included in the appendix the full textbook chapter on Linear Circuits, which may serve as a reminder or point of reference for those who wish to verify some details. This chapter is very well written (for whoever has learnt about Signals and Systems before), and I wouldn't want to spend a whole day simply rewriting word for word what's already written there. Again, if you don't know what Signals and Systems Theory is all about, you have two options: 1) Don't worry about it, you will be just fine with what I write about; 2) go read the first two chapters of the Alan Oppenheim textbook I mentioned above. If you choose 2), you have my respect, for whatever it's worth.

Here are things you should be comfortable with before starting to read through this chapter:

- Basics of electronic circuits, which is all reviewed in chapter 0
- Complex exponential basic mathematics.
- Architecture, function and application of the transconductance amplifier.

SOMETHING IMPORTANT IS MISSING IN THE CHAPTER, MAINLY THAT I AM NOT EXPLAINING WHY IT'S CALLED INTEGRATOR AND DIFFERENTIATORS.

6.1 Preliminary to Resistor Capacitor Circuits

This will be brief. I assume that most of the readers are reasonably comfortable with complex exponential already, and that Laplace Transforms are not unfamiliar either. Almost exclusively basing myself on the textbook.

⁴⁵ Here is a link to the PDF version: [https://eee.guc.edu.eg/Courses/Communications/COMM401%20Signal%20&%20System%20Theory/Alan%20V.%20Oppenheim,%20Alan%20S.%20Willsky,%20with%20S.%20Hamid-Signals%20and%20Systems-Prentice%20Hall%20\(1996\).pdf](https://eee.guc.edu.eg/Courses/Communications/COMM401%20Signal%20&%20System%20Theory/Alan%20V.%20Oppenheim,%20Alan%20S.%20Willsky,%20with%20S.%20Hamid-Signals%20and%20Systems-Prentice%20Hall%20(1996).pdf)

6.1.1 Complex Exponentials

All solutions to linear homogeneous equations are of the form e^{st} where s is a *complex number*.

$$s = \sigma + j\omega = M\cos(\phi) + jM\sin(\phi) \quad (114)$$

where $j = \sqrt{-1}$, σ is the real part of the complex number and ω is the imaginary part. M represents its *magnitude* and ϕ its phase. Magnitude and phase of a complex number obey the following relationships:

$$M = \sqrt{\sigma^2 + \omega^2} \quad (115)$$

$$\phi = \arctan\left(\frac{\omega}{\sigma}\right) \quad (116)$$

The magnitude of a complex number s is often denoted as $|s|$. Furthermore, applying the properties of complex exponentials, one can observe that:

$$e^{j\phi} = \cos(\phi) + j\sin(\phi) \quad (117)$$

$$e^{-j\phi} = \cos(\phi) - j\sin(\phi) \quad (118)$$

it follows that s can also be written as:

$$s = Me^{j\phi} \quad (119)$$

These notations can be used to solve higher order differential equations. As an example, we consider the second order linear homogeneous equation:

$$\frac{d^2}{dt^2}V + \alpha \frac{d}{dt}V + \beta V = 0 \quad (120)$$

Assuming e^{st} is an *eigenfunction*⁴⁶ and substitute for V :

$$s^2 e^{st} + \alpha s e^{st} + \beta e^{st} = 0 \quad (121)$$

Solving for s we obtain

$$s = \frac{-\alpha \pm \sqrt{\alpha^2 - 4\beta}}{2} \quad (122)$$

Consequently, if $\alpha^2 - 4\beta \geq 0$, s is real, otherwise it is complex.

6.2 Step and Delta function

If you do not know what this is, please see Textbook Linear Systems Chapter in appendix, section 8.3 and 8.4.

6.2.1 The Heaviside-Laplace Transform

By analyzing the example of the previous section we can make the following observation: Any time we substitute the eigenfunction e^{st} into a linear differential equation of order n , the following property obtains:

$$\frac{d^n}{dt^n} e^{st} = s^n e^{st} \quad (123)$$

In other words: We can consider s as an operator meaning *derivative* with respect to time. Similarly, we can view $\frac{1}{s}$ as the operator of *integration* with respect to time (Oliver Heaviside)

Formally, we write;

$$L\{y(t)\} = Y(s) = \int_{-\infty}^{\infty} y(t)e^{-st} dt \quad (124)$$

Notice that we write $y(t)$ for a signal in *time domain* and $Y(s)$ for a signal in *Laplace domain*. A signal in laplace domain is simply a signal in time domain which we have applied the Laplace transformation to (equation above). If you don't understand this, don't worry, figuring out how this works in practice is all that matter.

⁴⁶An eigenfunction is a nonzero solution of a second order linear homogenous differential equation

6.2.2 Transfer Function

Technically, to understand properly the concept of a transfer function, you should know what convolution, impulse response and Laplace transforms (more than what I introduced) are all about. But let's just make things very simple and say that we define the transfer function $H(s)$ as follows:

$$H(s) \equiv \frac{Y(s)}{X(s)} \quad (125)$$

which underlies very complex ideas just to say that it's the output divided by the input **both in Laplace domain** (see Figure 63).

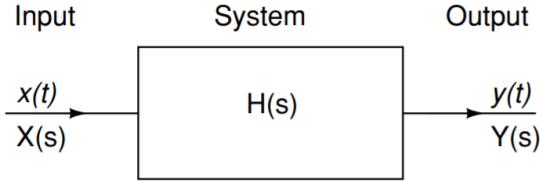


Figure 63: Typical black-box representation of a linear system. Its input is the signal $x(t)$ in time domain, and $X(s)$ in Laplace domain and its output is the signal $y(t)$ in time domain and $Y(s)$ in Laplace domain. Adapted from textbook

6.3 Resistor-Capacitor Circuits

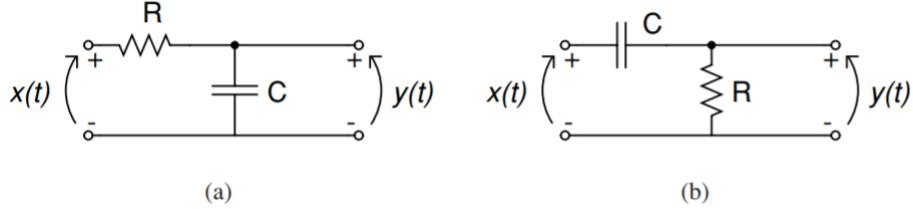


Figure 64: Resistor capacitor (RC) circuits. The signals $x(t)$ represent input voltages, and the signals $y(t)$ represent output voltages. (a) Integrator circuit (low pass filter); (b) Differentiator circuit (high pass filter). Adapted from textbook

I advise finding a memory hack to remember which is which. I know I begin with RC (not CR) which is low (you begin low and then you go high) and treats the "primitive" (integrator).

6.3.1 Solving Low pass Integrator RC circuit

Figure 64.a). Let's start by analyzing the circuit in a), and we need to start with KCL voltage law:

$$u_{in} = u_R + u_C, \quad (126)$$

where u_R and u_C are the voltage drops over the resistor R and capacitor C , respectively.

We have some current flowing in the branch and no current is flowing out of the main branch, so no current in $y(t)$. This is because, in electronics, we view open circuit as having infinite impedance. According to Ohm's law we have⁴⁷:

$$u_R = i \cdot R, \text{ and } u_C = \frac{i}{j\omega C} \quad (127)$$

⁴⁷I advise having a look again at section 0.1.9 about AC circuits and capacitance in complex domain if the following is not perfectly clear.

When combining these two equations, we get:

$$u_{in} = i \cdot R + i \cdot \frac{1}{j\omega C} = i \cdot (R + \frac{1}{j\omega C}) \quad (128)$$

In the case of this circuit, because voltage is the same across parallel branches, $u_{out} = u_c$, we can therefore write the transfer function $\frac{u_{out}}{u_{in}}$ as follows:

$$\frac{u_{out}}{u_{in}} = \frac{i \cdot 1/j\omega C}{i \cdot (R + 1/j\omega C)} = \frac{1}{j\omega RC + 1} \quad (129)$$

Now we can rearrange and essentially obtain a differential equation (remembering that $\frac{de^{j\omega t}}{dt} = j\omega e^{j\omega t}$)

$$j\omega \cdot RC u_{out} + u_{out} = u_{in} \equiv RC \frac{du_{out}}{dt} + u_{out} = u_{in} \quad (130)$$

You may wonder what this is all about, and why am I even doing this. I wondered myself, but don't worry, we'll get to it. It's all about transfer functions and understanding how the system output changes with respect to input change.

One important thing, RC has now appeared. As you can see, it is a constant (assuming resistance and capacitance are constant in a circuit) that scales the rate of change of the output with respect to time. This is called the time constant and is noted τ . This is critically important to understand charging and discharging time of capacitors (see Figure 65). The formal definition is: The circuit's time constant $\tau = RC$ is the time required to discharge the capacitor, through the resistor, to 36.8% (1/e) of its final steady state value.

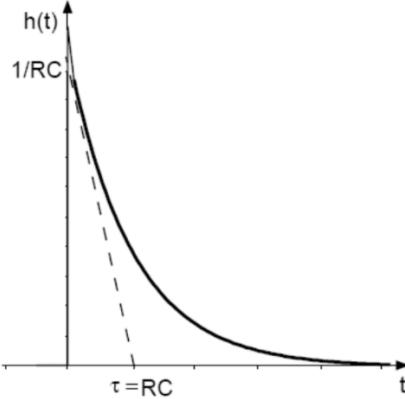


Figure 65: The circuit's time constant $\tau = RC$ is the time required to discharge the capacitor, through the resistor, to 36.8% (1/e) of its final steady state value. Adapted from lecture notes

$$H(s) = \frac{Y(s)}{X(s)} = \frac{1}{1 + RCj\omega} \quad (131)$$

Solving with Laplace transform In Laplace domain, the equation can be rearranged to:

$$\tau s Y(s) + Y(s) = X(s) \quad (132)$$

We can now also conveniently define the transfer function $H(s)$, taking $X(s) = 1$ (impulse):

$$H(s) = \frac{Y(s)}{X(s)} = \frac{1}{1 + \tau s} \quad (133)$$

6.3.2 Solving High pass Differentiator CR circuit

Figure 64.b). I won't be going into the derivation for this, but it really follows the same logic. In the end, we reach this differential equation:

$$j\omega \cdot \tau \cdot u_{out} + u_{out} = u_{in} \equiv \tau \frac{du_{out}}{dt} + u_{out} = \tau \frac{du_{in}}{dt} \quad (134)$$

Solving with Laplace transform In Laplace domain, the equation can be rearranged to:

$$\tau s Y(s) + Y(s) = s X(s) \quad (135)$$

Yielding the transfer function (also taking $X(s) = 1$):

$$H(s) = \frac{Y(s)}{X(s)} = \frac{\tau s}{1 + \tau s} \quad (136)$$

6.3.3 Frequency Domain Analysis

It may not be clear yet to you why are these circuits called lowpass/high pass filters, as well as differentiator/integrators. While we will look at the differentiator/integrator property in the next circuit, we can first start with some frequency domain analysis to make the filtering part clearer.

Let's now consider how this circuit responds to sinusoidal signals of different *frequencies*. Sinusoids have a very special relationship to shift-invariant linear systems, such as the one we are analyzing. When a sinusoidal signal is applied as input to a shift-invariant linear system, then its response will be another sinusoidal signal, with possibly a different amplitude and a different phase, but certainly with exactly the same frequency! That is, if the input is $x(t) = \sin(\omega t)$, the output will be $y(t) = A \sin(\omega t + \phi)$, where A and ϕ determine the scaling and shift.

Remember that the transfer function we obtained earlier for the RC low pass filter circuit was $H(s) = \frac{Y(s)}{X(s)} = \frac{1}{1 + \tau s}$. This is in Laplace domain, if we are playing with frequencies of signals, it makes sense to switch to the frequency (time) domain. In this domain $s = j\omega$ and the circuit's transfer function simply becomes:

$$H(j\omega) = \frac{1}{1 + j\omega\tau} \quad (137)$$

From this transfer function, we make two useful observations:

1. If the frequencies of the sinusoidal signals are small with respect to the circuit's time-constant ($\omega\tau \ll 1$), then the circuit's output will resemble its input ($Y(j\omega) \approx X(j\omega)$)
2. On the other hand, if the frequencies are large with respect to the circuit's time constant ($\omega\tau \gg 1$), then:

$$\frac{Y(j\omega)}{X(j\omega)} \approx \frac{1}{j\omega\tau} \quad (138)$$

These observations are also reflected in the plots of the transfer function's magnitude and phase (Figure 66). These plots are referred to as Bode plots and they are used to analyze the response of a dynamic system in terms of its transfer function. The magnitude of the transfer function is:

$$|H(j\omega)| = \frac{1}{\sqrt{1 + (\omega\tau)^2}} \quad (139)$$

and its phase is:

$$\phi = \arctan(-\omega\tau) \quad (140)$$

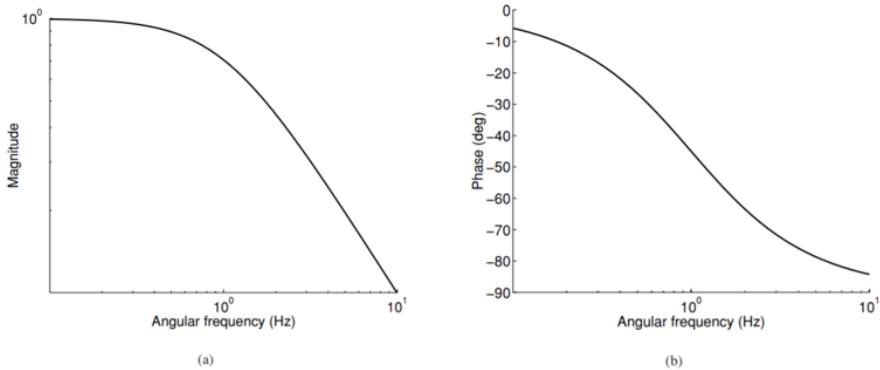


Figure 66: Bode plot of a first order linear system, such as the RC circuit of Fig. 64.a) Magnitude and b) Phase as a function of input signal angular frequency ω . Adapted from Textbook.

We define the frequency $\omega_{cutoff} = \frac{1}{RC} = \frac{1}{\tau}$ as the **cutoff frequency**. The cutoff frequency is the frequency at which, either above or below (depending on if you are using a low pass or high pass filter), the power output of a circuit is reduced to **1/2 of the passband power**⁴⁸. This is equivalent to a voltage (or amplitude) reduction of 70.7% of the passband, because voltage V^2 is proportional to power P. This happens to be close to -3 decibels and the cutoff frequency is frequency referred to as the -3dB point. In (Figure 66), it is set equal to 1.

Now let's look at a plot which really demonstrates how the response (output) of the input signal is reduced after filtering in figure 66.

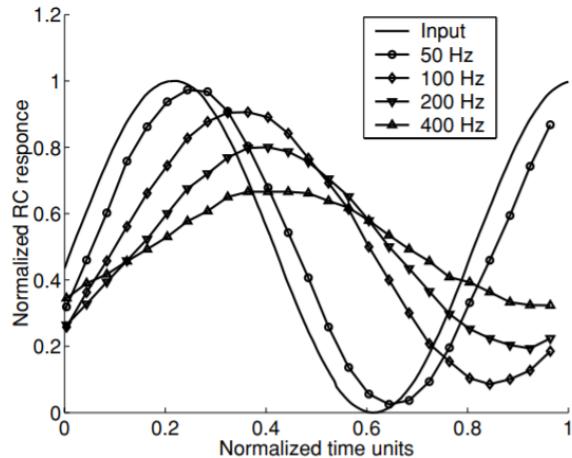


Figure 67: Response of an RC low-pass filter ($R = 10M\Omega, C = 1nF$) to input sinusoids of different frequencies. The input signals have been normalized to unity, and the outputs have been normalized with respect to the input. The time axis has also been normalized so that the responses to all the frequencies could be presented on the same graph. Adapted from Textbook.

The RC circuit of figure 64.a allows sinusoidal signals with frequencies lower than the cutoff frequency to pass virtually unchanged. On the other hand, the frequency components of the input signals that are above the cutoff frequency are attenuated. The phase lag between the input and the output of the system increases with ω (see Figure 66.b) and saturates at 90 degrees. Figure 67 shows experimental data measured from an RC lowpass filter with $R = 10M\Omega$ and $C = 1nF$. Sinusoids of increasing frequency were applied to the circuit and the corresponding responses were measured. To show the effect of a range of input frequencies on the circuit's response, all the data are plotted on a normalized scale. The responses have been normalized with respect to

⁴⁸Power, not amplitude. Remember that for an electrical circuit, power $P = \frac{V^2}{R}$

the input and time has been normalized to unity. As expected, the output signal is attenuated as the input frequency increases; and the phase lag between the input and output signals increases with increasing frequency.

6.3.4 Why are they called integrators and differentiators

There are physical explanations of why these are called integrators and differentiators, which should intuitively make sense. These unfortunately do not yet make sense to me. However, the mathematical intuition behind it is very elegant and simple to grasp, and perfectly explains why they bear these names.

Low Pass as Integrator: For a low pass filter, we established our transfer function as:

$$H(j\omega) = \frac{Y(j\omega)}{X(j\omega)} = \frac{u_{out}(t)}{u_{in}(t)} = \frac{1}{\tau j\omega + 1} \quad (141)$$

Now when we are beyond the cutoff frequency, that is: $\omega >> \omega_{cutoff} = \frac{1}{\tau}$, our transfer function becomes:

$$H(j\omega) = \frac{u_{out}(t)}{u_{in}(t)} \approx \frac{1}{\tau j\omega} \quad (142)$$

We can express u_{out} as a function of u_{in} , keeping in mind that u_{in} is a complex exponential: $u_{in} = e^{j\omega t}$, and its integral $\int u_{in}(t) dt = \int e^{j\omega t} dt = \frac{1}{j\omega} e^{j\omega t} = \frac{1}{j\omega} u_{in}(t)$

$$u_{out}(t) = \frac{1}{\tau j\omega} u_{in}(t) \propto \int u_{in} dt \quad (143)$$

And here is our integral!

High Pass as Differentiators: Following exactly the same logic, we note that at low frequencies, $\omega << \omega_{cutoff} = \frac{1}{\tau}$, our transfer function becomes:

$$H(j\omega) = \frac{Y(j\omega)}{X(j\omega)} = \frac{u_{out}(t)}{u_{in}(t)} \approx \frac{\tau j\omega}{1} \approx \tau j\omega \quad (144)$$

We can thus express u_{out} as a function of u_{in} and reach the elegant:

$$u_{out}(t) = \tau j\omega \cdot u_{in}(t) \propto \frac{\partial u_{in}(t)}{\partial t} \quad (145)$$

6.3.5 Summary about filters

- The figures we've just shown are for RC Low-Pass filters. The amplitude and power of the output signal decreases for high frequencies.
- The time constant $\tau = RC$ determines what high and low frequencies are, it also determines the cutoff frequency (at which output power is halved)
- Cutoff frequency $\omega_{cutoff} = \frac{1}{RC}$
- CR circuit (figure 64.b), is a high pass filter, which is just the same as we've shown but in reverse: it passes high frequencies and reduces low frequency.
- You can build a *band pass* filter from combining a high pass and low pass filter: the low pass will reduce all frequencies higher than your max desired frequency and the high pass will reduce all frequencies lower than your min desired frequency.
- A low pass filter acts as integrator for high frequencies, and high pass filter as differentiator in low frequencies/

- Why is this even useful? Imagine you're trying to process sound, like music for example, where you have some high frequency noise that is not your music (you barely hear it but it's there and alters your recording). Would be great if you could filter that out right? And only keep the frequencies of sound that are actually coming from the music (between 10 and 3000 Hz). This is just one dumb example on the top of my mind of why filtering is useful, and we'll see in the next section applications in Neuromorphic Engineering.

6.4 VLSI Integrators and Differentiators

Now we're going to blend the observations we've just made about low pass and high pass filters with what we know of the transconductance amplifier to do some cool things. Let's refresh our mind as to what the transconductance amplifier is all about:

- $I_{out} = I_1 - I_2 = I_b \frac{e^{\frac{\kappa V_1}{U_T}} - e^{\frac{\kappa V_2}{U_T}}}{e^{\frac{\kappa V_1}{U_T}} + e^{\frac{\kappa V_2}{U_T}}} = I_b \tanh\left(\frac{\kappa(V_1 - V_2)}{2U_T}\right)$
- $I_{out} \approx g_m(V_1 - V_2)$ with $g_m = \frac{I_b \kappa}{2U_T}$ for $V_1 - V_2 < 200mV$.
- $g_d = -\frac{\partial I_{out}}{\partial V_{out}} \approx \frac{I_b}{V_E}$
- $A = \frac{\partial V_{out}}{\partial(V_1 - V_2)} = \frac{\partial V_{out}}{\partial I_{out}} \frac{\partial I_{out}}{\partial(V_1 - V_2)} = \frac{g_m}{g_d} \approx \frac{\kappa V_E}{2U_T}$

What is the problem with the filters we've seen before? In VLSI technology, the time-constant of RC circuits implemented with passive elements cannot be changed once the chip has been fabricated. Both resistance and capacitance are fixed at design time by the geometries of the layout mask layers (this is the subject of NE2). By contrast, the transconductance amplifier has a transconductance that depends on its bias voltage which can be set on the operating chip. Therefore, this device can be used to design an integrator circuit that has an **adjustable time-constant**: This follower integrator, which was first designed by Carver Mead in 1989, is shown in Figure 69. But before looking at the follower integrator, let's look at the unity gain follower, which will allow us to better understand the follower integrator.

6.4.1 Unity Gain Follower

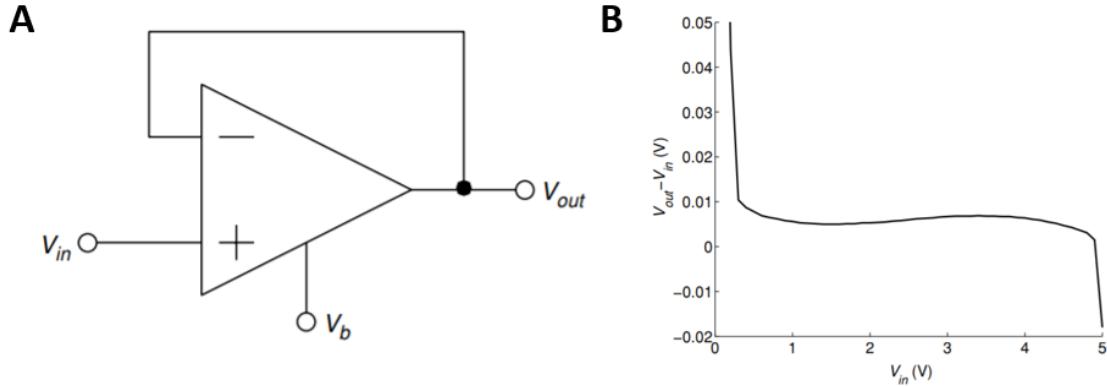


Figure 68: A) Unity Gain Follower circuit. B) Deviation of output voltage from input voltage for simple unity-gain follower. Adapted from Textbook.

Most circuit applications use the transconductance amplifier as part of a negative-feedback loop. The negative feedback ensures that the amplifier stays within its operating range. We mean by operating range the range where the difference between the input and output is small and the **conductance is linear**. The simplest negative feedback loop is obtained by short-circuiting the

output terminal and the inverting input terminal, as shown in Fig. 68.A. The transfer function of this circuit is given by:

$$\frac{dV_{out}}{dV_{in}} = \frac{A}{A+1} \approx 1 \quad (146)$$

Ok, how did this happen? Marc explained it to me, and it make sense. Let's go through it keeping in mind that $A = \frac{\partial V_{out}}{\partial(V_1 - V_2)}$:

$$\frac{\partial V_{out}}{\partial(V_{in} - V_{out})} = A \quad (147)$$

$$\partial V_{out} = A\partial(V_{in} - V_{out}) = A\partial V_{in} - A\partial V_{out} \quad (148)$$

Thus we reach back:

$$\frac{\partial V_{out}}{\partial V_{in}} = \frac{A}{A+1} \approx 1 \quad (149)$$

Now that this is clear, let's define define the input and output impedance:

$$\text{Input Impedance : } Z_{in} = \frac{dV_{in}}{dI_{in}} \rightarrow \infty \quad (150)$$

$$\text{Output Impedance : } Z_{out} = -\frac{dV_{out}}{dI_{out}} = \frac{1}{g_d} \approx \frac{V_E}{I_b} \quad (151)$$

Due to the large voltage gain A when it was open loop (when V_{out} was not an input), the transfer function is almost unity with $V_{out} \approx V_{in}$ now that we closed the loop. The circuit configuration is therefore called *unity-gain follower*. It is used as an impedance converter (also called a buffer) and it converts a high input impedance (because the Transconductance amplifier draws no current) into a lower output impedance. In contrast to the source follower presented in the simple circuit chapter, which is also used as an impedance converter, the unity-gain follower does not introduce a large voltage offset. The measured deviation of the output signal from the input signal, $V_{out} - V_{in}$, as a function of the input voltage is shown in Figure 68.B.

Long story short: it's the perfect resistor. Doesn't introduce gain, nor offset, and has adjustable conductance - let's make a filter out of it now!!

6.4.2 Follower Integrator

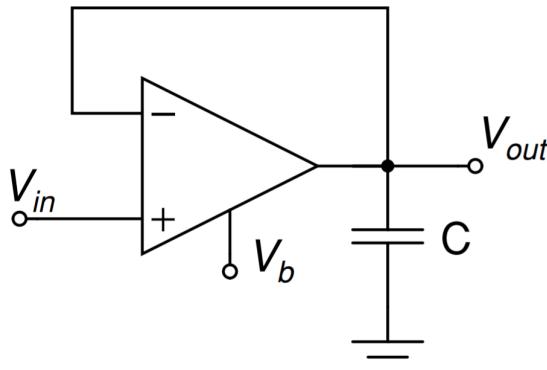


Figure 69: Follower Integrator circuit. The bias voltage V_b , which sets the transconductance amplifier's bias current I_b , can be used to modify the integrator's time-constant. Adapted from Lecture Notes. Adapted from textbook

This circuit is built out of a unity gain follower and a capacitor connected to the follower's output node. The input voltage is applied to the + terminal of the follower. If the circuit operates in subthreshold, we can apply Kirchoff's Current Law at the circuit's output node and reach the familiar looking equation:

$$C \frac{dV_{out}}{dt} = I_b \tanh\left(\frac{\kappa(V_{in} - V_{out})}{2U_T}\right) \quad (152)$$

Yes, it should be familiar, it looks a lot like what we did in the regular RC Integrator circuit, and because if we take conductance $g_m = \frac{\kappa I_b}{2U_T}$ (assuming small signal regime which is why we get rid of the tanh function), we can get back to the equation:

$$\frac{C}{g_m} \frac{dV_{out}}{dt} + V_{out} = V_{in} \quad (153)$$

Yes! This is exactly the differential equation that we derived for the low pass integrator filter in the previous section. Beautiful isn't it? Instead of a resistor, we used our transconductance amplifier **for which we can modulate the conductance**⁴⁹. So this circuit also has the following time constant: $\tau = C/g_m$. Note that the amplifier will operate in its linear range provided that V_{in} does not change too rapidly. Specially, we need $\frac{dV_{in}}{dt} < \frac{4U_T}{\tau}$. Under these conditions, the transfer function of our RC circuit and Follower Integrator are thus identical.

What happens when $\frac{dV_{in}}{dt}$ is large? When the AC component of V_{in} is large, the transconductance amplifier is no longer linear and the previous relation is invalid. For very large variations in V_{in} , the output current of the transconductance amplifier saturates at $\pm I_b$, which is the asymptote of the equation we first found ($C \frac{dV_{out}}{dt} = I_b \tanh\left(\frac{\kappa(V_{in} - V_{out})}{2U_T}\right)$). In this condition, the transconductance amplifier acts as a constant current source rather than as a linear conductor. While the difference $|V_{out} - V_{in}|$ is greater than $4UT$, V_{out} changes linearly with time. As the difference enters the small signal regime, the amplifier begins to behave as a linear conductor and V_{out} begins increasing (or decreasing) exponentially. Figure 70 shows the typical response of the follower-integrator circuit to a large step voltage input. The rate of change of output voltage $\frac{dV_{out}}{dt}$ in the region of constant slope, is defined as slew rate and is usually specified in units of $V/\mu s$. Slew rate is one measure of the performance limit of an operational amplifier, and is proportional to the maximum output current of the amplifier.

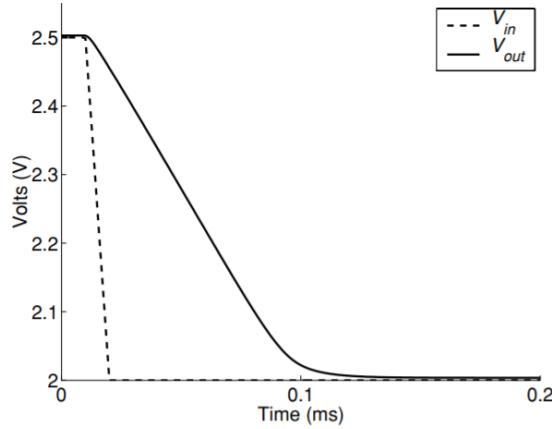


Figure 70: Large signal behavior of a follower-integrator. Response of the circuit to a large negative step input (dashed line). The output voltage V_{out} decreases linearly for large difference $V_{out} - V_{in}$ values and asymptotes exponentially for small differences. Adapted from textbook

⁴⁹By convention we speak of conductance for transconductance amplifier instead of resistance

6.4.3 Delay Lines

Cascade of RC integrators and follower-integrators

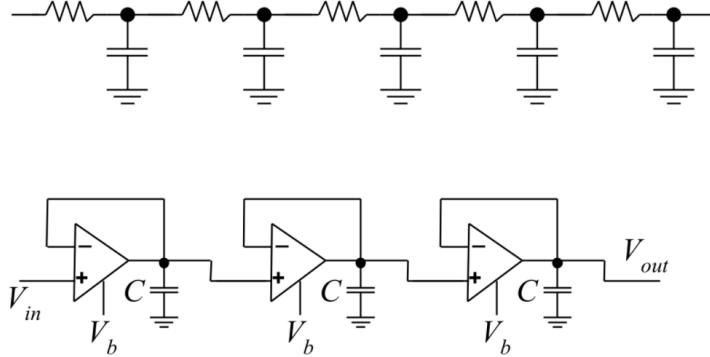


Figure 71: Cascade of RC integrators (top) and follower integrators (down). Adapted from textbook

We can compose (connect) multiple instances of the R-C integrator circuit, or multiple instances of the follower-integrator circuit (see Figure 71), in sequence to form a *delay line*. Each section in the delay line of **follower integrators** is modular (from a functional point of view) and independent of the other sections. The current out of the transconductance amplifier of one section can charge only the capacitor connected to the output node: It is not affected by the other sections connected to the output node. On the other hand, the sections of the R-C delay line are tightly coupled to one another. Current of one section flows into both the capacitor of that section and the resistor of the next. Because the characteristics of the single R-C circuit change when connected to other R-C circuits (as in Fig. 9.3), the transfer function of the composition of R-C circuits is not equal to the composition of individual transfer functions. Due to the modularity of the follower-integrator elements, the transfer function of the composition of follower-integrator circuits corresponds to the composition of their individual transfer functions. For example, if the number of elements in the delay line of follower integrators is n , we can simply use the transfer function of the regular RC integrator that was derived earlier and obtain:

$$\frac{V_{out}}{V_{in}} = \frac{1}{1 + \tau s}^n \quad (154)$$

This is a very useful property, because it allows the frequency response properties of the delay line to be evaluated analytically. Consider the case where sinusoidal inputs are applied to the circuit of Follower integrator delay line. The delay line's transfer function is:

$$\frac{V_{out}}{V_{in}} = \frac{1}{1 + j\omega \cdot \tau}^n \quad (155)$$

Exploiting the mathematical approximation that:

$$(1 + j\omega \cdot \tau) \approx (1 + \frac{1}{2}(\omega\tau)^2)e^{j\omega\tau} \text{ (for } \omega\tau \ll 1\text{)} \quad (156)$$

We can write the transfer function explicitly in terms of magnitude and phase:

$$\frac{V_{out}}{V_{in}} \approx \frac{1}{1 + \frac{n}{2}(\omega\tau)^2} e^{-j\omega n\tau} \quad (157)$$

where the pre-exponential ratio is the magnitude and the exponential's argument is the phase. From this equation, we can conclude that the magnitude of the output signal is attenuated by the factor $\frac{1}{2}(\omega\tau)^2$ as it crosses each section of the follower-integrator delay line, and the phase delay introduced by each section corresponds to $\omega\tau$ radians (equivalent to a time delay of τ). This

analysis is valid provided that each follower-integrator operates in its linear region. This is very useful to make better filters!

6.5 Laboratory : Integrator Circuits

Welcome to the time domain !

In this laboratory we study the follower-integrator circuit or FOI.

The FOI is special because its output follows the input at low frequencies, and integrates the input at higher frequencies. If you did some signal analysis before, the FOI is your low pass filter neuromorphic engineering edition.

We study the circuit in the time and frequency domain. For large and for small signals.

6.5.1 Time-domain response of small signal

The experiments starts at steady state. First we need to determine an input voltage that is high enough so that the transamp is capable of operating (i.e. 0.9V). More than that, the input should be corrected by the offset, which is a device specific operation.

At every step in the data collection phase we input this voltage to the circuit minus a small step input (i.e. 0.09V).

In small signals, the output increases exponentially with time based on the following equation :

$$V_{out}(t) = \Delta V_{in} \left(1 - e^{-\frac{t}{\tau}} \right) + V_{out}(t = 0).$$

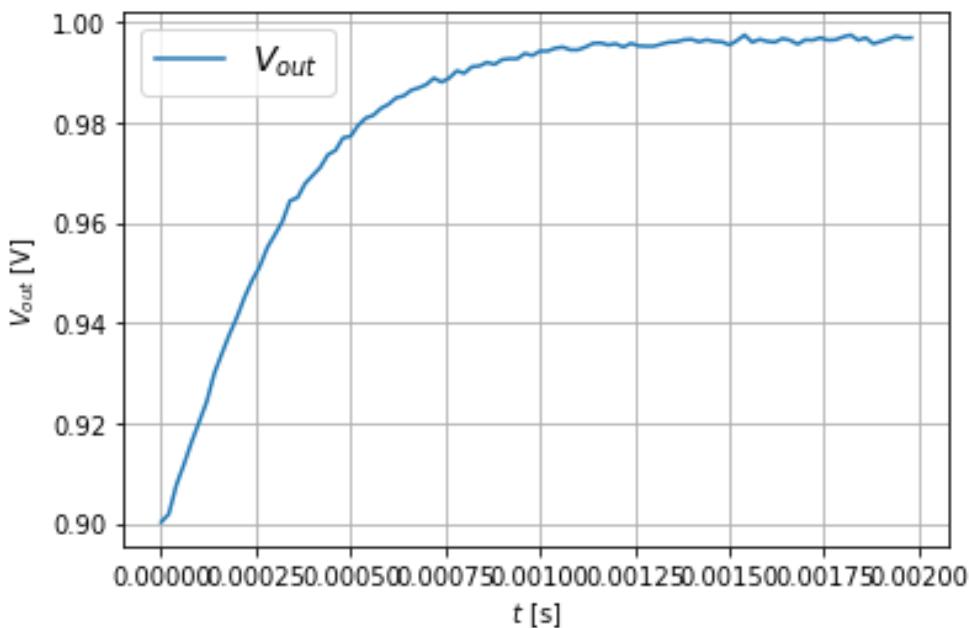


Figure 72: Step response of the follower-integrator for small signals.

To extract tau τ , we need to interpolate a function over our data and pass $V_{out}(t)$ to it.

It should be higher (i.e. 0.0003) than the estimates and the $V_{out}(t = 0)$ should be already greater than (i.e. 0.9V). We can recover the moment when the input step takes place by following the curve of V_{out} until it reaches 0V.

We found a kappa of around 0.8. for this experiment.

6.5.2 Time-domain response of large signal

The experiment time-domain response over large signals requires a large input step (i.e. 0.3V).

In this regime in the first part of the graph, $V_{out}(t)$ grows linear with time, therefore we can fit a line in this range.

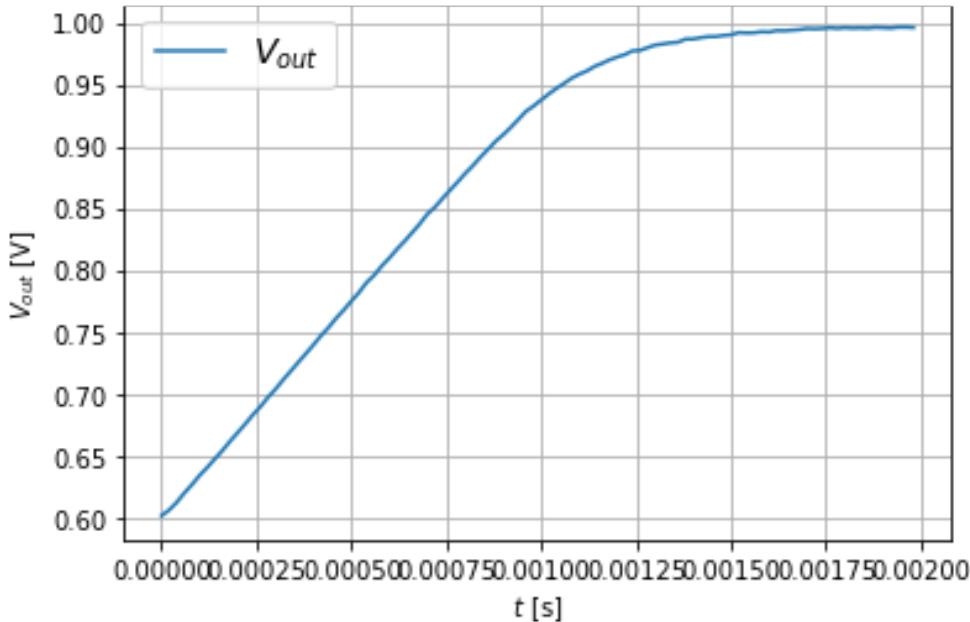


Figure 73: Step response of the follower-integrator for large signals.

The bias coefficient of the linear fitting gives us the slew rate (i.e 290). The bias current is found by multiplying the slew rate to the capacitance (i.e. 1e-12). We found a bias current 45pA higher than the theoretical expectation.

Tau and kappa in this experiment can be traced back to the linear part of the graph, therefore kappa does not have any significance.

6.5.3 Frequency-domain response

To measure the curve of the transfer function, we input a sin wave . The output of the FOI, follows the input current. However, it is also possible to change the regime to integrator. The output current of the latter will be translated to the left with respect to the input.

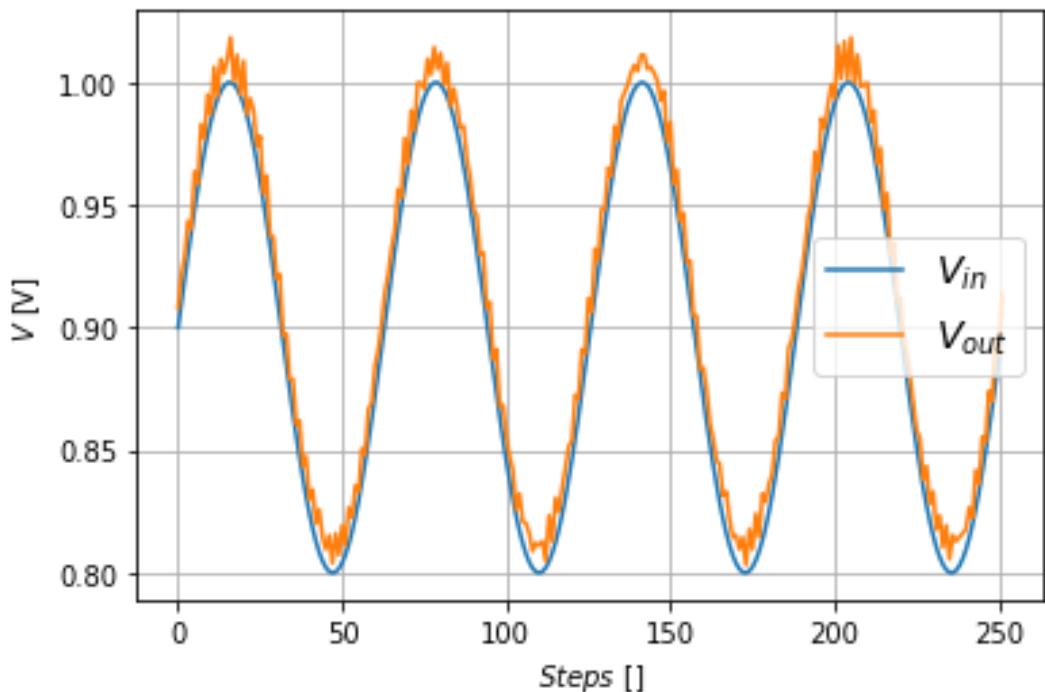


Figure 74: Follower Integrator, Wave Form

6.6 Things you should know

Here is a list of things you should know about this chapter for the exam:

- How to compute the time-constant of a low-pass filter and how to estimate it from measurements.
- How to change the time constant of a follower integrator circuit.
- What it means to use a lowpass filter as integrator and highpass filter as differentiator
- Why the differentiator and integrator are not actually real but only an approximation over some frequencies defined by the time constant.
- How to sketch the transfer function of a differentiator circuit, showing the time constant on the sketch.
- How to implement a simple follower-differentiator
- How these circuits behave when driven with large signals

7 Current Mode and Winner Take All

Current mode happens to be a relatively new thing in circuit design, and a process of its own. This is particularly true in analog design where historically, information is *represented by voltage* at nodes of circuits. In current mode, it's just the opposite. Without entering the details of the reason for such a switch which are complicated and beyond the scope of this module, it is found that current-mode circuits can operate at low power-supply voltages and over a wide range of currents. Their advantages include higher bandwidth, higher dynamic range, and they are more amenable to lower power supplies. There are whole textbooks that have been written on only the topic of CMOS current mode operation. Thanks to current mode circuits, we really have managed to shrink power supply: in the 90s we're talking 5V power supply for a chip, today you have state of the art implementation with a few milivolts as power supply, which is not a lot. There are now a wide range of different circuits both in academia and industry using current mode circuits.

Long story short, we switch from having inputs, outputs and parameters as voltages to having them as currents, and it happens to be better as it needs less power and some other advantages that I sadly not fully understand yet.

What you should be comfortable with before starting to read through this chapter:

- Forward and Reverse Current from Ohmic and Saturated subthreshold MOSFET.
- Early Voltage

7.1 Translinear Circuits

7.1.1 Short intro

Before we start on explaining the translinear principle, let's touchbase on what translinear means altogether. Trans can mean a few different things in its Latin origin, but here, it means "going beyond". So translinear kinda means that we're going beyond linearity. But linearity of what? Linearity of current-voltage characteristic. Yes, our course is all about operating transistors in subthreshold and making them translinear: we have an exponential relationship between current and voltage, and not a linear one. That's it, that's what this fancy terms is all about.

So a translinear circuit is a circuit that uses translinear circuit elements: MOSFETs operating in subthreshold (or BJT) and hence in the exponential current voltage relation.

7.1.2 Translinear principle

Static tranlinear circuits were invented by the late Barrie Gilbert, and there is actually a whole lecture on Youtube where he explains what this is all about⁵⁰. Let's write the formal definition: "*In a closed loop containing an even number of forwardbiased junctions arranged so that there are an equal number of clockwise-facing and counter clock-wise facing polarities, the product of the current densities in the clockwise direction is equal to the product o the current densities in the counter clock-wise direction.*". This actually is best visualized by looking at a simple circuit (as opposed to the weird one shown in lecture notes):

⁵⁰<https://www.youtube.com/watch?v=LQNJVtcFrCc>

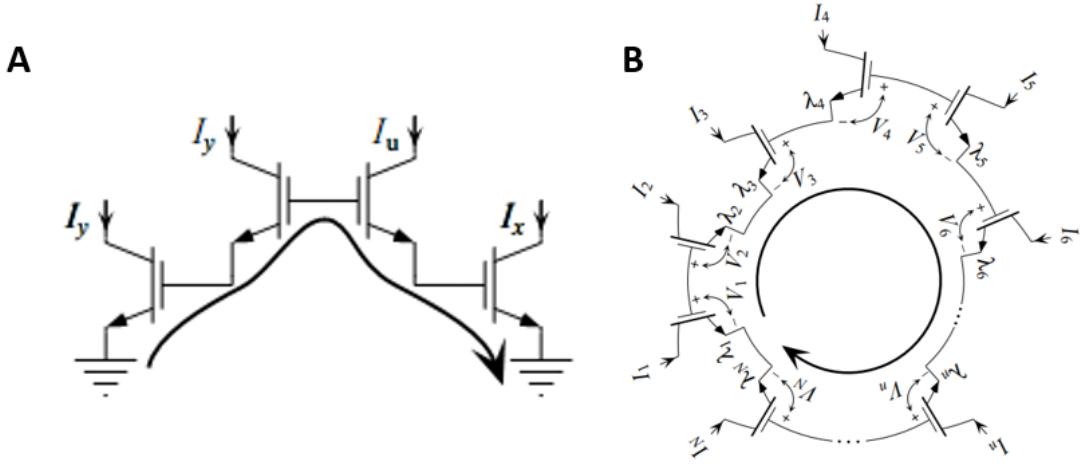


Figure 75: Two circuits with a translinear loop. A) Simple Translinear circuit. Adapted from Wikipedia. B) Complicated translinear circuit. Adapted from Lecture notes.

Consider figure 75.A, you can see that there is an equal amount of clockwise facing and counter clockwise facing polarities BJTs⁵¹. Remember Kirchoff's current law? (*In any complete loop within a circuit, the sum of all voltages across components which supply electrical energy (such as cells or generators) must equal the sum of all voltages across the other components in the same loop.*). We can apply that to figure 75.A (and B as well for the matter, and let's just focus on B because of the notation) and observe that the voltage around the loop that goes from V_1 to V_N must be 0. In other words, the voltage drops must equal the voltage increases. We call this a translinear loop. Mathematically, this becomes

$$\sum_{k=1}^N V_{Fk} = 0 \quad (158)$$

If you notice on the transistors, the V_{Fk} on figure 75.B are the gate to source voltages of the transistors. Imagining the image drew MOSFETs and not BJTs, that all transistors operate in saturation, and that $\kappa = 1$, we have each $V_{Fk} = U_T \ln\left(\frac{I_{ds_k}}{I_0}\right)$. So now we can rewrite the previous sum as follows:

$$\sum_{k=1}^N U_T \ln\left(\frac{I_{ds_k}}{I_0}\right) = 0 \quad (159)$$

We can exploit the logarithm property that $\ln(x \cdot y) = \ln(x) + \ln(y)$ and that $\ln(1) = 0$ we reach the final:

$$\prod_{k=1}^N \frac{I_{ds_k}}{I_0} = 1 \quad (160)$$

The reason why this is useful will be visible soon enough. But mostly, if we have a circuit with a loop where sum of voltages is 0, we'll be able to apply this principle and we'll be happy.

7.2 The Current Conveyor

7.2.1 Introduction

In voltage-mode circuits, the main building block used to add, subtract, amplify, attenuate, and filter voltage signals is the operational amplifier. In current-mode circuits, the analogous building block is the *current conveyor*. It really originates from a need to have a current mode equivalent

⁵¹Yes, the images are shown with BJTs rather than MOSFETs. Don't worry too much about it.

of the op-amp for the current mode function - which is funny because the idea of current conveyor was subsequently invented before the circuit was invented (discovered/engineered or whatever you want to call it).

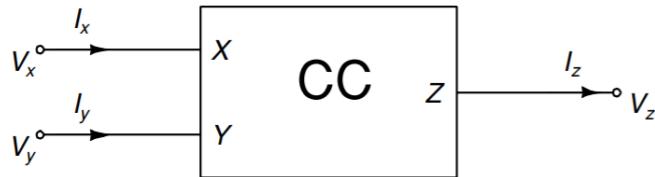


Figure 76: The current conveyor representation (not circuit). Adapted from Textbook.

If we want to do current mode computation, we need a building block that can convey current from input terminal to output terminal while *decoupling* the circuit, just like an op-amp where V_1 and V_2 are the input which come with infinite impedance, and then you have the output V_{out} - so you completely decouple the input from the output. We need the same in current mode. We also want to use this building block to do basic computations (add, multiply, filter etc...). Here is the formal definition:

- The potential at the output terminal (Z) is independent of the current applied at node Y .
- An input current that is forced into node X results in an equal amount of current flowing into node Y .
- The input current flowing into node X is conveyed to node Z , which has the characteristics of a high output impedance current source.

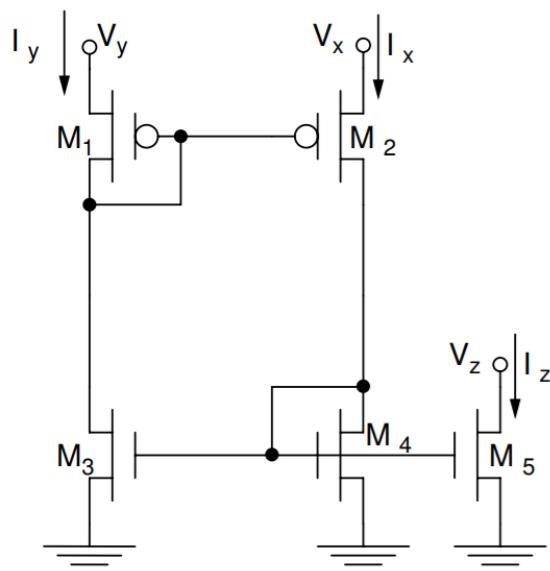


Figure 77: A circuit which satisfies the current conveyor definition. Adapted from Lecture notes.

Briefly looking at Figure 77, which is a potential current conveyor, the input terminals are X and Y , the output is at Z . If you apply a current through Y , it results in the same current through X because of the current mirror, which is copied again at $M3$ and $M4$ which yields that whatever flows through Y is conveyed to Z . Again this is just one example of a current conveyor, which actually is just unity in this circuit.

7.2.2 Basic subthreshold current conveyor

Figure 78 shows a current conveyor that is commonly used at the Institute of Neuroinformatics, and the one we should know about.

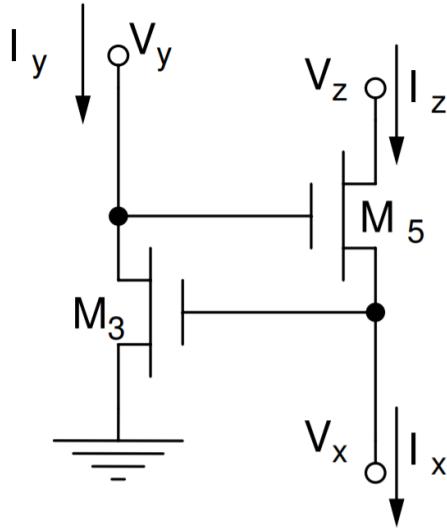


Figure 78: A more common current conveyor circuit. This is the one you should know for the exam. Adapted from Lecture notes.

Here we force a current into node V_y which is copied into node V_x . How does that work? Let's go through it:

- If you *force* the current I_y through M_3 , the gate voltage of M_3 *must adapt* to the log of the current flowing through it⁵². So $V_x \propto \ln(I_y)$.
- Now we set I_x (which is another input of the circuit) to be equal to I_y (or proportionality factor of I_y), we will exactly like before force the $V_{gs} = V_y - V_x$ of M_5 to follow $\log(I_x)$.
- This then results into I_z which must follow I_x
- So to summarize, two take aways about this: 1) $I_z = I_x$ and 2) $V_x \propto \ln(I_y)$
- It decouples input current from output current, i.e. while the input and the output current are equal, the voltage at the output is independent of the input current.

It's used in a broad range of circuits such as:

- Low pass filters
- Multiplier circuits
- Winner take all circuits
- Silicon Neurons
- Current-mode silicon retinas

⁵²because (omitting κ, I_0 and U_T for simplicity: $I = \exp(V_{gs})$ so $V_{gs} = \log(I)$)

7.3 The current conveyor as a multiplier

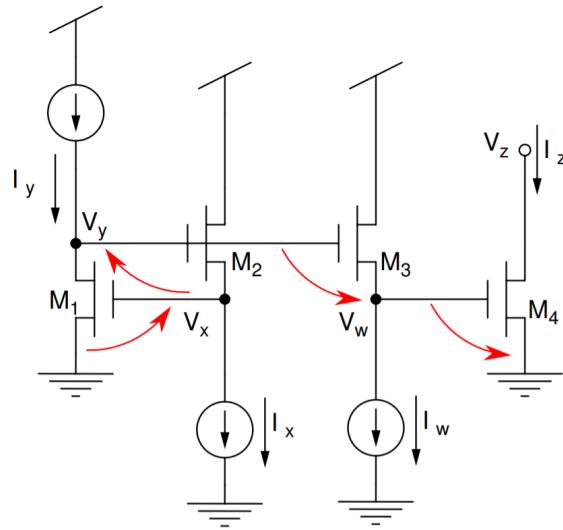


Figure 79: Current conveyor as a multiplier. Adapted from Lecture notes.

Let's look at what this circuit does. The first thing to notice is that building block (on the left) is exactly the current conveyor we've seen before, with I_x , I_y and I_z etc. Now, let's look at the V_{gs} 's, remembering the translinear principle. Look at the red arrows representing the V_{gs} of each transistor: two of them (on the left) go *from* ground and the two others (on the right) go *into* ground. Yes! The translinear principle applies. Let's put that formally:

- First $V_{gs_1} = V_x - 0$
- Second $V_{gs_2} = V_y - V_x$
- Third $V_{gs_3} = V_y - V_w$
- Fourth $V_{gs_4} = V_w - 0$

Applying the translinear principle, this yields:

$$(V_x - 0) + (V_y - V_x) - (V_y - V_w) - (V_w - 0) = 0 \quad (161)$$

$$\rightarrow V_{gs_1} + V_{gs_2} - V_{gs_3} - V_{gs_4} = 0 \quad (162)$$

$$\rightarrow I_x \cdot I_y \cdot I_w^{-1} \cdot I_z^{-1} = 0 \quad (163)$$

$$\text{Yielding the elegant : } I_z = \frac{I_x \cdot I_y}{I_w} \quad (164)$$

One must remember that this is an approximation as we assumed in the translinear principle derivation earlier that $\kappa = 1$, which is not true in practice. As a reminder, κ is usually 0.7-0.8 in practice.

7.4 The Gilbert Normalizer

I think this circuit is not to be known for the exam but it's way too elegant not to look at. Let's keep it quick:

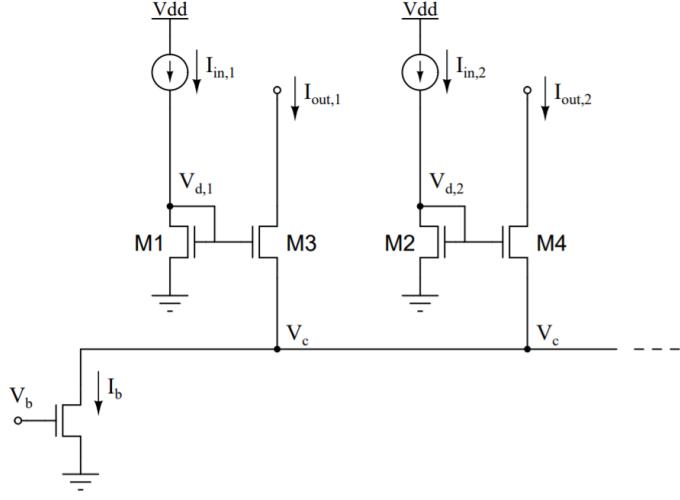


Figure 80: Gilbert Normalizer circuit. Adapted from Lecture notes.

We assume subthreshold and saturation in all transistors:

Because of Kirchoff's current law, the sum of currents flowing through each branch $I_{out,j}$ must be equal to I_b : Let's now note some things about I_{in} and I_{out} relationship, which is just driven by a current mirror arrangement.

$$I_{in_i} = I_0 e^{\frac{\kappa V_{di}}{U_T}} \quad (165)$$

$$I_{out_i} = I_0 e^{\frac{\kappa V_{di}}{U_T} - \frac{\kappa V_c}{U_T}} = I_{in_i} e^{-\frac{\kappa V_c}{U_T}} \quad (166)$$

$$\text{Since } I_b = \sum_j I_{out_j} \quad (167)$$

By using the nice trick of multiplying equation 166 by $\frac{I_b}{I_b} = \frac{I_b}{\sum_j I_{out_j}} = 1$, we reach:

$$I_{out_j} = \frac{I_b}{I_b} I_{in_j} e^{-\kappa V_c / U_T} = \frac{I_b}{\sum_j I_{out_j}} I_{in_j} e^{-\kappa V_c / U_T} = \frac{I_b I_{in_j} e^{-\kappa V_c / U_T}}{e^{-\kappa V_c / U_T} \sum_j I_{in_j}} \quad (168)$$

$$\text{Thus reaching : } I_{out_j} = I_b \frac{I_{in_j}}{\sum_j I_{in_j}} \quad (169)$$

Hence implementing some kind of normalization of all I_{out_j} flowing. Pretty cool!

7.5 Winner Take All circuit

A winner-take-all (WTA) circuit is a network of competing cells (neural, software, or hardware) that reports only the response of the cell that has the strongest activation while suppressing the responses of all other cells. The circuit essentially implements a *max()* function. These networks have been implemented in analog VLSI and applied to a wide variety of tasks, including selective attention, auditory localization, visual stereopsis, smooth pursuit/tracking, detection of heading direction and more. Today, it's a circuit that is basically everywhere.

What's the biological basis behind Winner Take All? It really is believed that populations of neurons do some kind of selective amplification to the broad range of input signals that they receive. Imagine you're at a loud cocktail party and there is a lot of noise but you suddenly hear your name being said, your brain will instantly cancel out everything else but the origin of the sound that you recognized as your name (localization and pitch/frequency of the voice). This is just one example. Anyways, in general, these circuits are typically used to implement and model competitive mechanisms among population of neurons. In this section we will analyze a class of

WTA networks that emulate biological networks, consisting of a cluster of excitatory neurons that innervate a global feedback inhibitory neuron.

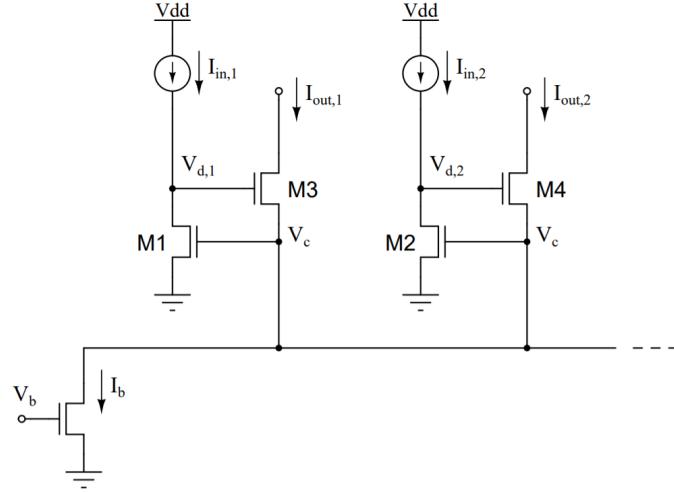


Figure 81: Current Mode Winner Take All. Adapted from Lecture.

The circuit of Fig. 81 is a continuous time, analog circuit that implements a WTA network. It processes all the (continuous-time) input signals in parallel, using only two transistors per input cell, and one global transistor that is common to all cells. Collective computation and global connectivity is obtained using one single node common to all cells. As can be seen in Fig. 81, each cell is built out of a current conveyor. The WTA network is modular and can be extended to N cells, by connecting additional cells to the node V_c . Input currents are applied to the network through current sources which are implemented for example using subthreshold pFETs. There are three conditions that determine the functioning of this WTA circuit, and we'll look at each of them individually.

7.5.1 $I_{in_1} = I_{in_2} = I_{in}$

You should intuitively see (also from the previous normalizer circuit) that when $I_{in_1} = I_{in_2}$, $I_{out_1} = I_{out_2} = I_b/2$. Let's see how that works formally. First, from the current conveyor, we established that $V_c \propto \ln(I_{in})$, which if written formally gives:

$$V_c = \frac{U_T}{\kappa} \ln\left(\frac{I_{in_1}}{I_0}\right) \quad (170)$$

Because $I_{in_1} = I_{in_2}$, it follows that $V_{d1} = V_{d2}$, which yields that $I_{out_1} = I_{out_2} = I_b/2$. As simple as that.

To remember:

- $V_c = \frac{U_t}{\kappa} \ln\left(\frac{I_{in}}{I_0}\right)$
- $I_{out_1} = I_{out_2} = I_b$
- $V_{d1} = V_{d2} \approx V_c + V_b$

7.5.2 $I_{in_1} \gg I_{in_2}$

This is a bit trickier. We first need to recall from Chapter 2 and 3 that the subthreshold current flowing through a transistor can be divided into a *forward* component, I_f and a *reverse* component, I_r . When the transistor's source voltage V_s is approximately equal to its drain voltage V_d (so when we are in the Ohmic region), I_r becomes comparable to I_f . Now let's just keep this property in mind and look at the circuit for the case where $I_{in_1} \gg I_{in_2}$.

When $I_{in_1} \gg I_{in_2}$, if the drain voltage of M_1 is in saturation, the dominant component of its drain current will be in the forward direction and its gate voltage V_c will increase such that $I_{d1} = I_{f1} = I_0 e^{\kappa V_c / U_T} = I_{in_1}$. Although the two input currents I_{in_1} and I_{in_2} are different, the forward component of the drain currents of M_1 and M_2 are equal ($I_{f1} = I_{f2}$) because the two transistors have a common gate voltage V_c , and both their sources are tied to ground. The drain current I_{d2} of transistor M_2 can only be equal to the input current I_{in_2} under the following conditions:

$$I_{f2} - I_{r2} = I_{in_2} \quad (171)$$

$$\text{which implies that : } I_{r2} = I_{f2} - I_{in_2} \quad (172)$$

$$\text{which implies that : } I_{r2} = I_{in_1} - I_{in_2} \gg 0 \quad (173)$$

The reverse component of I_{d2} becomes significant only if V_{d2} decreases enough for M_2 to operate in its ohmic region. In this case, the output transistor M_4 is effectively switched off, and $I_{out_2} = 0$. Consequently, M_3 sources all the bias current ($I_{out_1} = I_b$), with V_{d1} satisfying the equation $I_b = I_0 e^{\kappa V_{d1} - V_c}$.

To remember:

- $V_c = \frac{U_t}{\kappa} \ln\left(\frac{I_{in_1}}{I_0}\right)$
- $I_{out_1} = I_b; I_{out_2} = 0$
- $V_{d1} = V_c + V_b$
- $V_{d2} \approx 0$

Typical δV that make this condition apply are actually as small as 5mV. It's only when δV is smaller than this that we are in the operating regime described below.

7.5.3 $I_{in_1} = I_{in_2} \pm \delta I_{in}$

To analyze the circuit in this regime, we must consider the Early effect of the transistor operating in the saturation region. Recall that considering the Early Effect, current in a transistor in saturation is:

$$I_{ds} = I_{sat}\left(1 + \frac{V_{ds}}{V_e}\right) \quad (174)$$

where V_e is the early voltage.

Assume that the two input currents I_{in_1} and I_{in_2} are initially equal. In this case, the transistors M_1 and M_2 operate in saturation region: the output voltages V_{d1} and V_{d2} will settle to a common value and the output currents I_{out_1} and I_{out_2} are both equal to $I_b/2$ as established previously. If we now increase the input current I_{in_1} by a small amount δI and apply the previous equation to transistor M_1 , then the drain voltage V_{d1} increases by:

$$\delta V = \frac{\delta I}{I_{sat}} V_e \quad (175)$$

As V_{d1} is also the gate voltage of transistor M_3 , the I_{out_1} will be amplified by an amount proportional to $e^{\delta V}$. The constraint of Kirchoff's current law imposes that I_{out_2} decreases by the same amount in steady state. This reduction means that gate voltage V_{d2} of M_4 must decrease by δV .

The gain of the competition mechanism $\frac{\delta V}{I}$ in the small signal regime is directly proportional to the Early voltage and inversely proportional to I_{sat} . The Early voltage depends on the geometry of the transistors and is fixed at design time. On the other hand I_{sat} depends on V_c , which changes with the amplitude of the input currents.

To remember:

- $I_{ds} = I_{sat}\left(1 + \frac{V_{ds}}{V_e}\right)$
- $V_{d2} = V_{d1} - V_e \frac{\delta I}{I_{sat}}$
- $I_{out_2} < I_{out_1}$

7.5.4 Experimental data

We can see experimental data on figure 82.

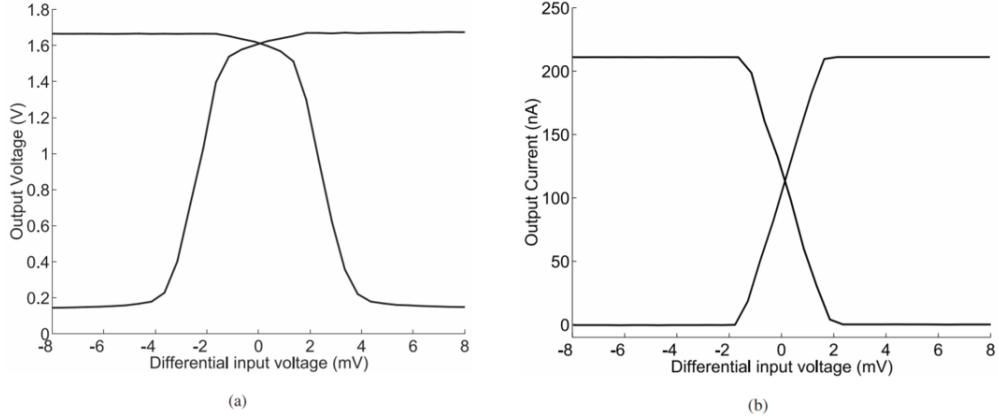


Figure 82: Responses of the current conveyor two-cell WTA circuit. a) Voltage output (V_{d1} and V_{d2}) versus the differential input voltage. b) Current output (I_{out1} and I_{out2}). The bias voltage $V_b = 0.7V$. The small difference in the maximum output currents is due to device mismatch effects in the read-out transistors of the two cells. Adapted from textbook.

We can see the output voltages (V_{d1} and V_{d2}) and output currents (I_{out1} and I_{out2}) of the circuit, in response to the differential input voltage δV which encodes the ratio of the input currents that were provided by pFETs operating in subthreshold. V_{in1} was set to 4.3V while the gate voltage of the input current pFET V_{in2} was set to $V_{in2} = V_{in1} + \delta V$, thereby allowing to test all different three case scenarios by changing δV . We can clearly see that experimental result go well with our theoretical findings!

7.6 Things you should know

Here is a list of things you should know about this chapter for the exam:

- What's the idea behind current mode circuits?
- What are the 3 principles of a current conveyor.
- Draw the basic current conveyor, and work through its function.
- Draw and explain the function of the Gilbert Normalizer.
- How does the WTA circuit work.
- Can you reason through its behavior?
- How does the bias current affect the performance of the WTA?

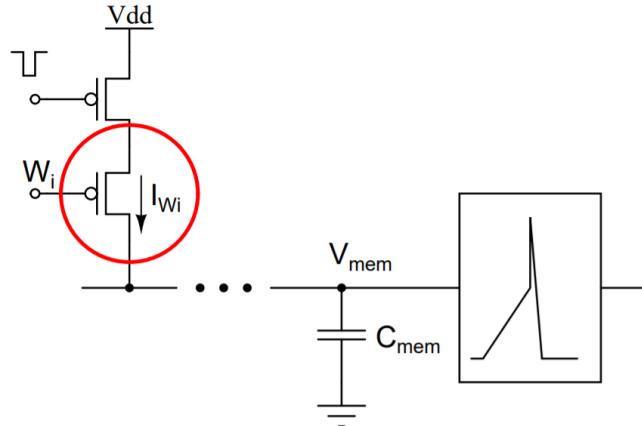


Figure 83: Simple circuit of a VLSI synapse in a pulse-based neural network.

8 Silicon Synapses

So far, we have seen several circuits that demonstrate fairly complex functionality by using only a small number of circuit elements. But what does this all have to do with Neuromorphic Engineering? Didn't we want to emulate the behaviour of neural systems? It turns out that by combining the previously introduced circuits we are actually able to create circuits that have very similar properties to real cortical synapses and neurons. In this chapter, we look into VLSI synapses. Make sure that you are familiar with chapter 0.2 and 0.3.1 where we briefly introduced neurons and synapses as well as the Perceptron, a very simple artificial neuron model. In the Perceptron model, the synapse implements the multiplication between the neuron's input signal X_i and its corresponding weight W_i . Silicon synapses in "classical" neural networks similarly only act as multipliers. However, we want to emulate neural systems and are therefore interested in a more biologically plausible synaptic behaviour. Biological systems communicate with spikes, i.e. short activation "pulses". So how can we generate a similar weighted pulse with our silicon synapse?

8.1 VLSI Synapses in Pulse-based Neural Networks

Figure 83 demonstrates the simplest circuit to generate an output pulse. It consists of two pFET transistors and one capacitor. The pFET MOSFET on the top receives an input in the form of a gate voltage change. By decreasing the gate voltage, a current is generated. The strength of the current is shaped by the lower pFET whose gate voltage W_i we can adjust. The generated current I_{wi} finally charges the circuit's capacitor and the charging and the discharging of the capacitor creates our desired voltage "spike" as shown on the right. Based on the capacitor equation $C \frac{dV}{dt} = I$, we can describe the voltage change as follows:

$$\Delta V_i = \frac{I_{wi}}{C_{mem}} \Delta t \quad (176)$$

The above equation demonstrates that the strength of the output spike can be changed by either changing the bias voltage W_i which shapes the current I_{wi} or by changing the duration Δt of the input. The demonstrated circuit allows us to implement an excitatory synapse. By replacing the pFETs with two nFETs (connected to ground instead of V_{dd}), the generated current I_{wi} flows into the opposite direction. The capacitor is consequently discharged first and we get an inhibitory synapse. As we are using a current source - our I_{wi} - to (dis)charge the capacitor, the voltage changes linearly. This is however not biologically plausible. In biological systems, a synapse acts as a linear integrator. If you recall chapter 6.3, an integrator is equivalent to a low-pass filter. Its impulse response is a decaying exponential. Remember that in subthreshold, there is an exponential relationship between the voltage and the current, e.g. as shown in the equation of a saturated pFET in subthreshold:

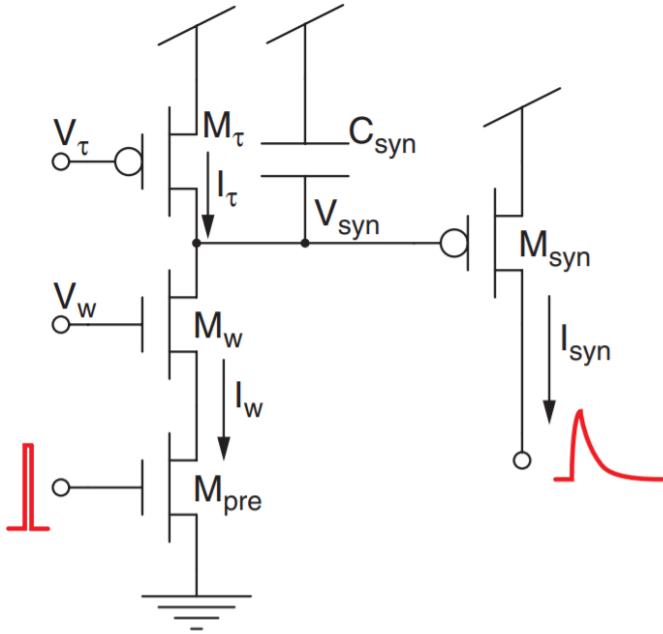


Figure 84: Exponentially decaying integrator circuit.

$$I(t) = I_0 e^{\frac{\kappa}{U_T} (V_{dd} - V_g(t))} \quad (177)$$

If we linearly change the gate voltage, we get an exponential change of current. So let's use this intrinsic property of subthreshold MOSFETs to build a biologically plausible synapse.

8.2 Exponentially decaying integrator circuit

Figure 84 shows the circuit of an exponentially decaying integrator that exploits the exponential behaviour of subthreshold MOSFETs and shows similar behaviour to a synapse. We assume that all transistors operate in subthreshold and in saturation. Consequently, we get the following equations for the circuit's currents:

$$I_w = I_0 e^{\frac{\kappa V_w}{U_T}} \quad (178)$$

$$I_\tau = I_0 e^{\frac{\kappa (V_{dd} - V_\tau)}{U_T}} \quad (179)$$

$$I_c = C \frac{d}{dt} (V_{dd} - V_{syn}) \quad (180)$$

$$I_{syn} = I_0 e^{\frac{\kappa (V_{dd} - V_{syn})}{U_T}} \quad (181)$$

Let's go through the circuit's behaviour step by step. We assume that initially the capacitor is fully charged, so $V_{syn} = V_{dd}$. As V_{syn} is also the gate voltage of the pFET M_{syn} , this means that there is no current I_{syn} flowing. When we get a pulse input, i.e. a positive gate voltage at transistor M_{pre} , we generate a current flow. The strength of the current is dependent on the gate voltage of the above transistor M_w and we therefore denote the generated weighted current as I_w . You might wonder why we use two transistors to generate our weighted input. By separating our weight factor from the input pulse, we are able to process digital inputs that are either completely on or completely off. Additionally, we can introduce dynamic behaviour to our weighting factor V_w . This will be introduced in more detail in section 8.4.1. We can assume that our generated input current I_w is always significantly larger than the current I_τ . Consequently, we have a constant current flowing out of the capacitor, i.e. discharging it, and a linear decrease of V_{syn} .

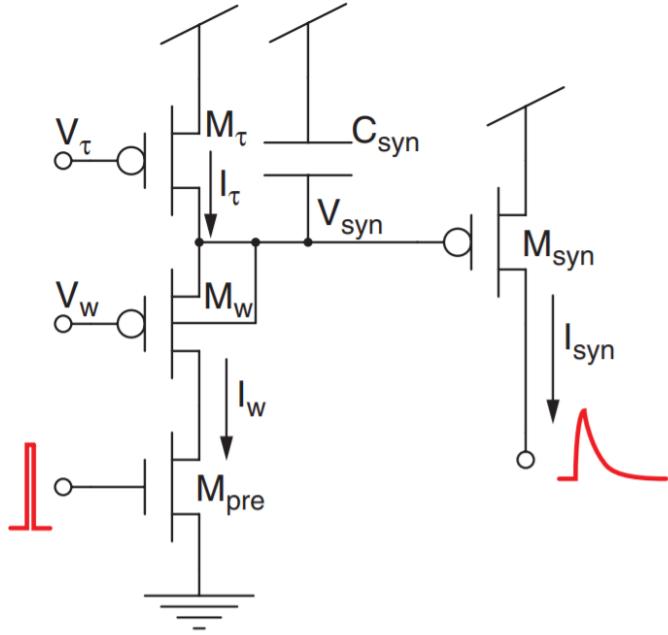


Figure 85: Log-domain pulse circuit.

This decrease leads to $V_{gs} > 0$ in the pFET M_{syn} and we get an output current I_{syn} . When there is no pulse input at M_{pre} anymore, no current I_w is flowing. The current I_τ now flows into the capacitor and linearly charges V_{syn} until it reaches V_{dd} again. Due to the exponential relationship, the linear increase of V_{syn} causes an exponential decrease of I_{syn} until the output is completely shut off. The charge and discharge phase of the circuit can be described by the following equations:

$$I_{syn}(t) = \begin{cases} I_{syn}^- e^{+\frac{(t-t_i^-)}{\tau_c}} & \text{(charge phase)} \\ I_{syn}^+ e^{-\frac{(t-t_i^+)}{\tau_d}} & \text{(discharge phase)} \end{cases} \quad (182)$$

We can rewrite the equation for I_{syn} as follows:

$$I_{syn}(t) = I_0 e^{n\Delta t(\frac{1}{\tau_c} + \frac{1}{\tau_d}) - \frac{t}{\tau_d}} = I_0 e^{-\frac{\tau_c - f\Delta t(\tau_c + \tau_d)}{\tau_c \tau_d} t} \quad (183)$$

where n is the number of input spikes and $f = \frac{n}{t}$ the spike frequency. Note that for a large spike frequency, the capacitor does not have the time to charge up again. The voltage V_{syn} eventually saturates at 0 and the resulting output current I_{syn} becomes constant and independent of the input spikes. Looking at the above equation of I_{syn} , this is the case when the exponent becomes a positive value and eventually "explodes". We therefore get the following condition for our spike frequency:

$$f\Delta t(\tau_c + \tau_d) > \tau_d \implies f < \left(\frac{\tau_c}{\tau_c + \tau_d}\right) \frac{1}{\Delta t} \quad (184)$$

8.3 Log-domain Pulse Integrator

A similar circuit with synaptic behaviour is the log-domain pulse integrator. It is shown in figure 85. The only difference compared to the previous circuit is that the transistor M_w is replaced by a pFET and its n-well is now connected to the source voltage V_{syn} . Why do we do that? Remember from chapter 3.9.4 that transistors are usually distorted by the body effect. By connecting the n-well to V_{syn} , we can cancel out this effect and thereby ensure that the thresholds of the M_w and M_τ transistors remain similar. The updated equation of the weighted input current is:

$$I_w = I_0 e^{\frac{\kappa}{U_T} (V_{syn} - V_w)} \quad (185)$$

Let's try to find an analytical description for our output current I_{syn} . First, we want to know how our current changes over time.

$$\frac{d}{dt} I_{syn} = -I_{syn} \frac{\kappa}{U_T} \frac{d}{dt} V_{syn} \quad (186)$$

$$\text{With } \tau = \frac{C_{syn} U_T}{\kappa I_{syn}} : \frac{d}{dt} I_{syn} = -I_{syn} \frac{\kappa}{U_T} \frac{d}{dt} V_{syn} \quad (187)$$

$$\text{With } I_c = I_\tau - I_w : \tau \frac{d}{dt} I_{syn} = -I_{syn} + I_{syn} \frac{I_w}{I_\tau} \quad (188)$$

We can further rewrite our input current I_w as follows:

$$I_w = I_0 e^{-\frac{\kappa}{U_T} (V_w - V_{syn})} = I_0 e^{-\frac{\kappa}{U_T} (V_w - V_{dd})} e^{\frac{\kappa}{U_T} (V_{syn} - V_{dd})} = I_{w0} \frac{I_0}{I_{syn}} \quad (189)$$

where I_{w0} is the current I_w at the beginning of the charge phase, i.e. when $V_{syn} = V_{dd}$. By using this new equation of I_w , we can get the following differential equation for our output current I_{syn} :

$$\frac{d}{dt} I_{syn} + I_{syn} = \frac{I_0 I_{w0}}{I_\tau} \quad (190)$$

While this circuit provides a good approximation of synaptic behaviour, it highly restricts our control of the output current I_{syn} . Both I_0 and I_τ are fixed values. Additionally, we have to ensure that I_w remains in subthreshold (I w has to be large because the fixed factor is small). So how can we construct a more controllable silicon synapse?

8.4 Diff-Pair Integrator (DPI) Synapse

The diff-pair integrator (DPI) synapse is visualized in figure 86. It is named after its main component, a differential pair circuit. In order to understand its behaviour, we will analyze its complementary circuit shown in figure 87 instead. The circuit in figure has the advantage that we can apply the translinear principle to it. The principle states that all gate-to-source voltages within a closed loop add up to zero. Note that the arrows always show from the source to the drain of the transistor. We get the following sum of voltages:

$$(V_g - 0) - (V_g - V_s) + (V_c - V_s) - V_c = 0 \quad (191)$$

As we have seen in chapter 7.1.2, we can transform the above sum to a product of currents. Note that when we subtract the gate voltage of a pFET, we multiply the **positive** current of the corresponding transistor. We therefore get:

$$I_{th} I_1 I_2^{-1} I_{out}^{-1} = 1 \implies I_{th} I_1 = I_2 I_{out} \quad (192)$$

Because of Kirchhoff's current law, we know that: $I_{in} = I_1 + I_2$ and $I_2 = I_\tau + I_c$. We replace the currents in above equation and get:

$$I_{th} (I_{in} - I_\tau - I_c) = (I_\tau + I_c) I_{out} \quad (193)$$

Now we would like to find an alternative equation for I_c that is independent of the voltage.

$$I_{out} = I_0 e^{\frac{\kappa V_c}{U_T}} \quad (194)$$

$$\frac{d}{dt} I_{out} = I_{out} \frac{\kappa}{U_T} \frac{d}{dt} V_c \implies \frac{d}{dt} V_c = \frac{d}{dt} I_{out} \frac{U_T}{\kappa I_{out}} \quad (195)$$

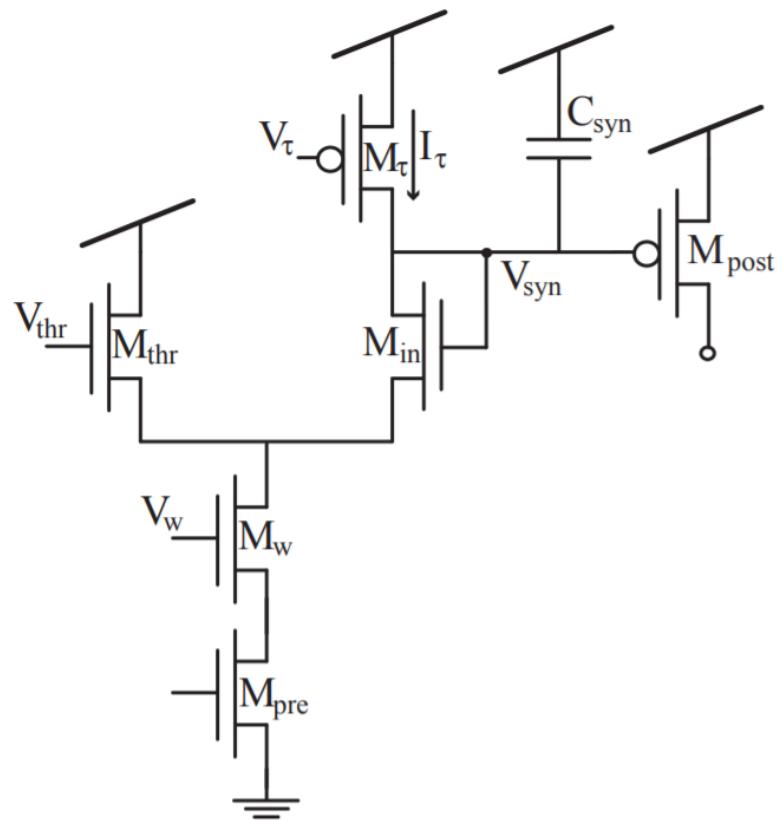


Figure 86: VLSI circuit of a diff-pair integrator (DPI) synapse.

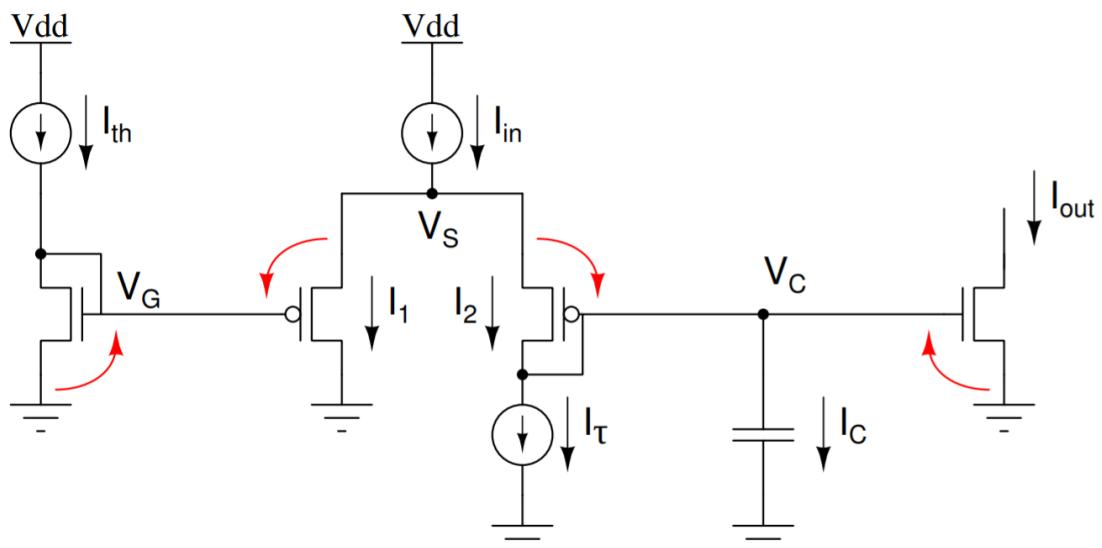


Figure 87: Complementary VLSI circuit of the DPI synapse.

$$I_c = C \frac{d}{dt} V_c \implies I_c = C \frac{U_T}{\kappa I_{out}} \frac{d}{dt} I_{out} \quad (196)$$

By inserting this equation into (193) and rearranging, we get the following nonlinear differential equation for our output current:

$$\tau \left(1 + \frac{I_{th}}{I_{out}}\right) \frac{d}{dt} I_{out} + I_{out} = \frac{I_{th} I_{in}}{I_\tau} - I_{th} \quad (197)$$

where $\tau = \frac{C U_T}{\kappa I_\tau}$. So what happens now when we get a positive input current I_{in} ? As soon as I_{in} becomes larger than I_τ , the capacitor is being charged. This leads to a linear increase of V_c , assuming a constant current input. Due to the exponential relationship between gate voltage and current, I_{out} consequently starts to increase exponentially. As I_{th} is a constant, the term $\frac{I_{th}}{I_{out}}$ in the above equation eventually becomes zero when we have an input current. Similarly, the term I_{th} on the right hand side of the equation becomes negligible for $I_{in} \gg I_\tau$ as $\frac{I_{in}}{I_\tau} \gg 1$. We can therefore simplify (197) to the following linear differential equation:

$$\text{If } I_{in} \gg I_\tau : \tau \frac{d}{dt} I_{out} + I_{out} = \frac{I_{th}}{I_\tau} I_{in} \quad (198)$$

Unlike the previously introduced circuits in section 8.2 and 8.3, the behaviour of our synapse is now dependent on a variable current, I_{th} . This allows us to easily control our circuit's output.

There are two important things to note about the derived circuit. For simplicity, we previously assumed that $\kappa = 1$. However, in reality κ is usually around 0.7. In the DPI synapse, it turns out that all κ values cancel each other out. We therefore only have to ensure that all κ values are equal, in particular $\kappa_N = \kappa_P$ but we do not have to assume that $\kappa = 1$ anymore. Consequently, our derived equations very precisely model the circuit's actual behaviour. Another observation we make is that our calculations exploit the linear range of operation of the circuit's differential pair. We therefore have to ensure that V_g and V_s are within a similar range, i.e. that $|V_g - V_s|$ is small.

Let's have a look at our DPI synapse in figure 86 again. Based on the same derivations, its behaviour can be modelled by the same linear differential equation:

$$\text{If } I_w \gg I_\tau : \tau \frac{d}{dt} I_{syn} + I_{syn} = \frac{I_{thr}}{I_\tau} I_{th} \quad (199)$$

where $\tau = \frac{C U_T}{\kappa I_\tau}$. The only difference is that our input current I_{in} is represented by the weighted current I_w . Our resulting circuit provides a biologically plausible and controllable model of a cortical synapse.

8.4.1 Short-term Depression

In biological systems, the response of a synapse decreases when it receives input spikes at a high rate. This phenomenon is called short-term depression. It is visualized in figure 88. At the bottom of the figure, we see the firing pattern of the presynaptic neuron. As shown at the top of the figure, the synaptic information the postsynaptic neuron receives continuously decreases with every incoming spike. By using two different transistors M_{pre} and M_w to separate our input pulse from its weight, we are actually able to implement this behaviour by replacing V_w with an additional circuit. This circuit is shown in figure 89.

The bottom transistor receives the same input pulse as our transistor M_{pre} in the DPI synapse. Whenever we receive a pulse, i.e. an increase of the gate voltage, we generate a current I_r . This current flows out of the capacitor C_d , consequently discharging it and reducing the capacitor's voltage V_w . This is the voltage we use as the weighting gate voltage in our DPI synapse. The more pulses we receive, the more current flows out of the capacitor and the lower V_w becomes. When we don't receive any input pulses, the diode V_a eventually charges up the capacitor again and we retrieve our initial weighting voltage V_w .

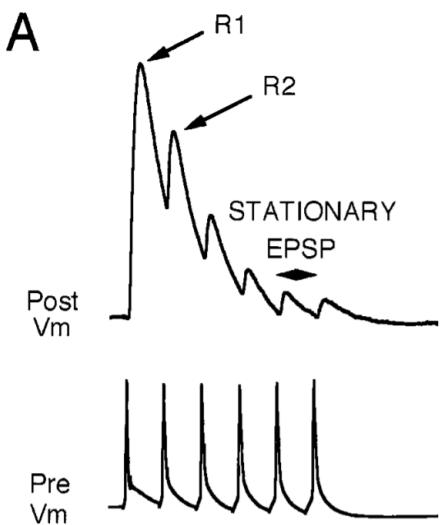


Figure 88: Membrane voltage of a pre- and postsynaptic neuron that undergo short-term depression.

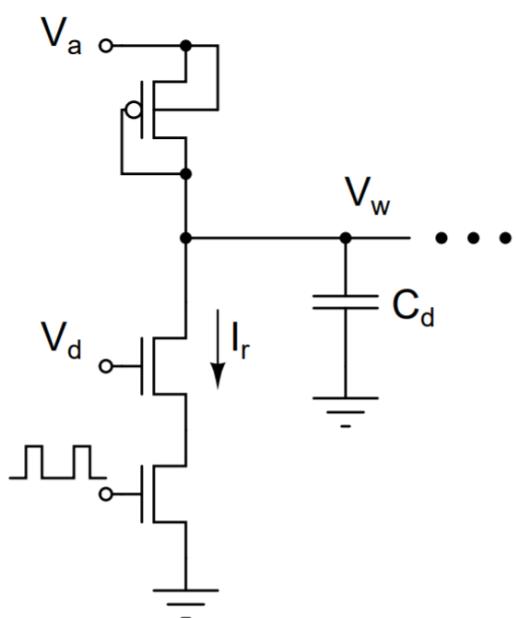


Figure 89: VLSI circuit that implements short-term depression.

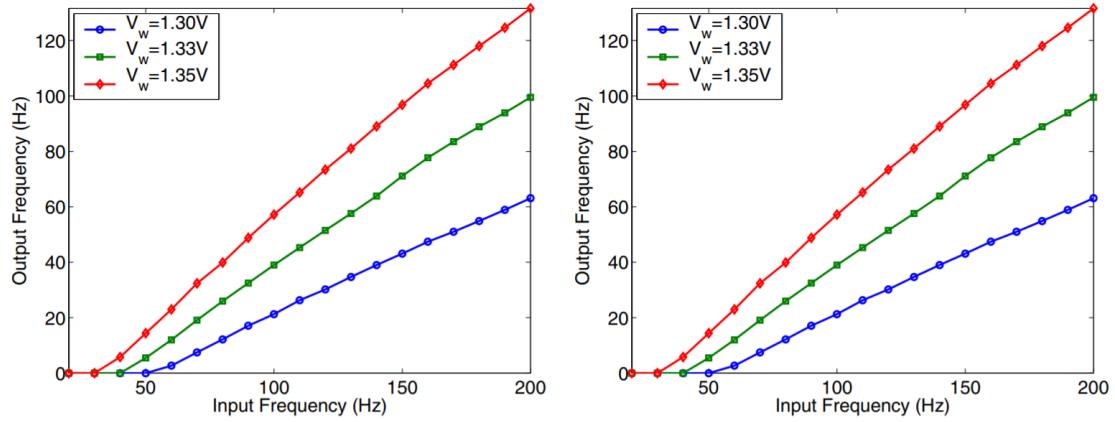


Figure 90: The response of the DPI synapse to a spike train (left) and the Rectified Linear Unit (ReLU) function (right).

There is one very interesting thing to note when comparing the frequency of our input pulses with the frequency of our output spikes. This relationship is shown in ???. First of all, the impact of our weight onto our output becomes clear. As expected, the higher our weight, the higher our output frequency. But more importantly - does the shape of the graph look familiar to you? Exactly, it is a ReLU function! For those unfamiliar with it, the ReLU (Rectified Linear Unit) function is an activation function that is very commonly used in neural networks. Its output is zero for any negative input or the positive input itself. Its shape is shown in figure ???. So why don't we simply use a ReLU function instead of our complex synapse circuit? When we are only interested in applying a ReLU activation onto our inputs, this is indeed the better way. The introduced synapse circuits, however, allow us to incorporate more complex dynamic behaviour that also occurs in biological synapses.

8.5 Laboratory :Silicon Synapses

In this session, we study how synaptic circuits generate current when stimulated by voltage pulses. Specifically we will measure the response of a synapse to a single pulse, and to a sequence of spikes.

To this extend, we will measure the response properties of the diff-pair integrator (DPI) synapse

8.5.1 DPI synapse

The DPI synapse receives a voltage pulse train, V_{pulse} , as input and outputs a corresponding synaptic current, I_{syn} and voltage V_{syn} . Bias parameters V_{weight} V_{tau} affect the amplitude and decay of the response, while V_{thr} acts as an additional weight bias. C_{syn} sizing was chosen for a capacitance of 2pF.

We need to tune these parameters and observe the behavior of the DPI synapse. To test the basic response of the circuit we send 2 input pulses and sample 10 points, with a delta t of 0.02 between the samples.

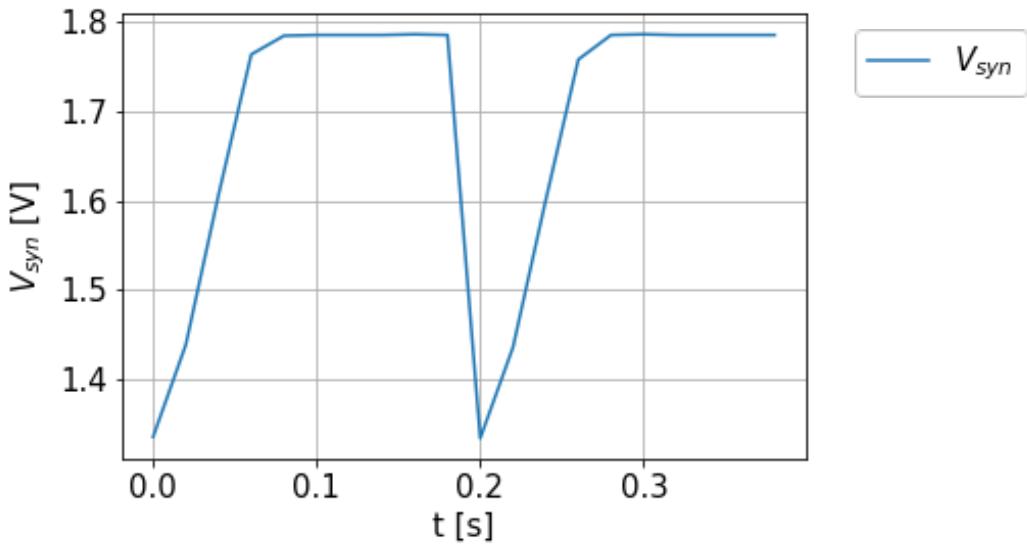


Figure 91: Measured values of V_{syn} as a function of time

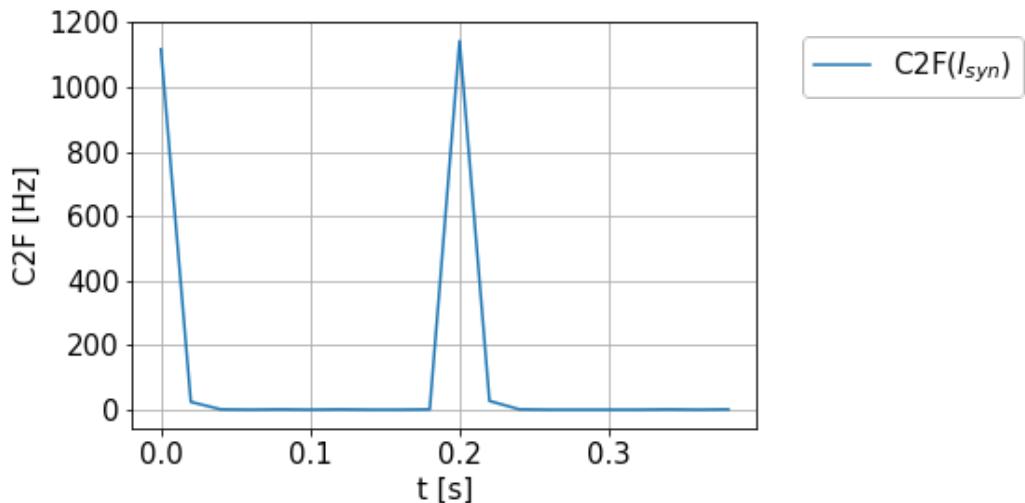


Figure 92: Measured C2F values of I_{syn} as a function of time

We found that if we increase V_{weight} or V_{thr} the amplitude of I_{syn} increases and V_{syn} will need more time to recharge and return back to it's initial state (1.8V).

With a larger V_{tau} or V_{pulse} the opposite is truth. The amplitude of I_{syn} shrinks and V_{syn} returns to the pre-spike state much faster.

8.6 Test Yourself

You should be able to answer the following questions for the exam (mainly taken from the winter study sheet).

- The schematic for a synapse circuit.
- How the synaptic current changes as a function of the presynaptic frequency and the synaptic weight.

- How the firing frequency of an adaptive neuron changes as a function of the presynaptic frequency and the biases to the adapting synapse.
- Which differential equation describes the output current of the DPI synapse? Under which condition does it hold?
- What are the advantages of using two separate transistors for the input pulse?
- How can we implement short-term depression?
- What is the disadvantage of the Log-domain pulse integrator?

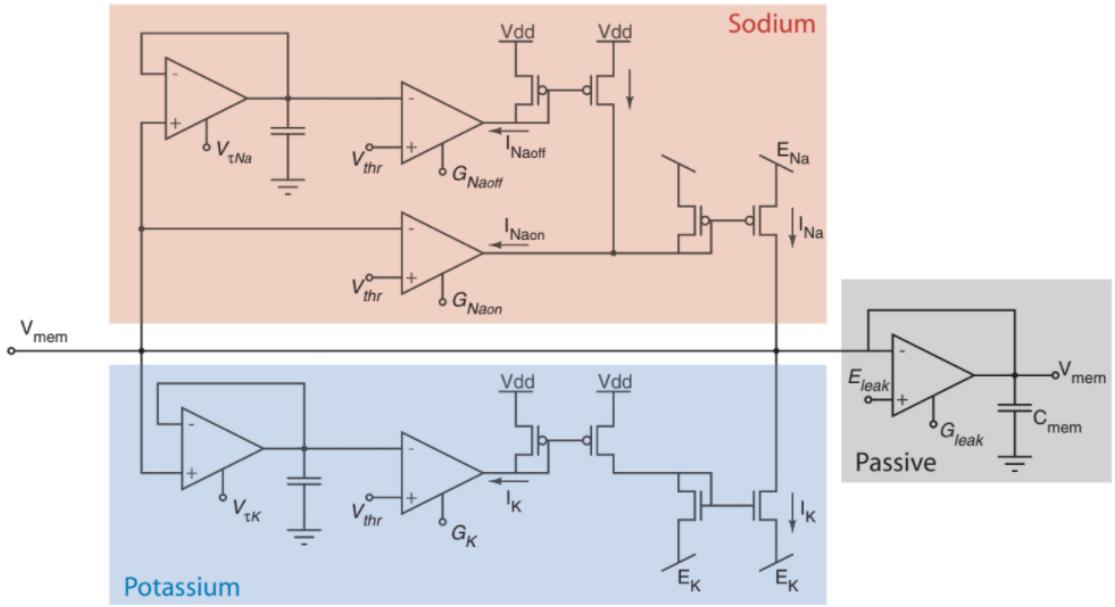


Figure 93: Conductance-based silicon neuron.

9 Silicon Neurons

In the previous chapter, we introduced a biologically plausible circuit that models the behaviour of a cortical synapse. Synapses are responsible for transmitting information between neurons. Neurons themselves act as integrator circuits that integrate over their inputs and generate an action potential once their integrated input crosses a threshold. Figure 11 in section 0.2 visualized the typical course of an action potential. Let's have a look how we can model an equivalent behaviour with our VLSI circuits.

9.1 Conductance-based silicon neuron

In 1991 Misha Mahowald and Rodney Douglas developed a conductance-based silicon neuron that demonstrates remarkably similar properties to those of cortical neurons. The proposed circuit is shown in figure 93. It consists of three part - the passive leak current of the cell membrane, the positive feedback of an action potential that is regulated by sodium in cortical neurons and the negative feedback that is regulated by potassium.

Passive component

Let's have a look at the circuit's passive component at first. It consists of a follower integrator with output voltage V_{mem} and input voltage E_{leak} . The input voltage represents the circuit's resting potential. Remember that the behaviour of the follower integrator can be described by the following equation:

$$C \frac{d}{dt} V_{out} = I_b \tanh\left(\frac{\kappa(V_{in} - V_{out})}{2U_T}\right) \quad (200)$$

The change in output voltage is dependent on the capacitor's capacitance and the output current of the transconductance amplifier. If $V_{mem} = E_{leak}$, we don't have any output current and consequently the capacitor is neither charged nor discharged. Our output voltage V_{mem} remains unchanged. If V_{mem} becomes larger than E_{leak} , a negative current is generated. This current discharges the capacitor and V_{mem} decreases until it equals E_{leak} again. On the other side, when V_{mem} is smaller than E_{leak} , we get a positive current that charges the capacitor until $V_{mem} = E_{leak}$. Just looking at our passive component, the membrane voltage always get pulled

back to the resting potential E_{leak} .

Sodium positive feedback component

The circuit's positive feedback component consists of another follower integrator and two transconductance amplifiers. Remember from section 5.2.3, that the transconductance amplifier acts as a comparator if it is operated in voltage mode. V_{out} converges towards ground if the input V_+ is larger than V_- and towards V_{dd} if $V_- > V_+$. If the circuit's membrane voltage V_{mem} is smaller than V_{thr} , we get a large output voltage V_{out} in the lower transconductance amplifier of the sodium component. This voltage is connected to the gate voltage of a pFET current mirror and we consequently have no output current I_{Na} . Once our membrane voltage is increased by a positive input signal, V_{mem} eventually becomes larger than V_{thr} and the output voltage V_{out} starts to converge towards zero. This decrease of the gate voltage generates a current flow in the pFET current mirror and we get a positive current I_{Na} . This current is equivalent to the current that pushes into our transconductance amplifier as defined by its equation $I_{out} = I_b \tanh(\frac{\kappa(V_{thr} - V_{mem})}{2U_T})$, denoted as $I_{Na, on}$. Note that the source of our current mirror is not set to V_{dd} but to E_{Na} . Like this we can ensure that our current is bound by E_{Na} . As the connection between I_{Na} and the membrane voltage V_{mem} acts as a capacitor, V_{mem} increases which leads to an even larger current I_{Na} . This is the circuit's positive feedback loop. At the same time, V_{mem} is the input voltage of the component's follower integrator. Once V_{mem} starts to increase, the output voltage adapts to the input voltage with a time delay. This delay is dependent on the integrator's capacitance and its bias voltage, denoted as $V_{\tau Na}$ as described in section 6.4.2. Once the output voltage becomes larger than the threshold voltage V_{thr} of the top transconductance amplifier, the output voltage converges towards zero and we generate a current flow at our pFET current mirror. This generated current, denoted as $I_{Na, off}$, is connected to the output current of our previous transconductance amplifier. For $I_{Na, off} = I_{Na, on}$, all the current that pushes into our transamp is pulled from $I_{Na, off}$. Consequently, there is no input current into our low current mirror anymore and I_{Na} becomes zero. The top transconductance amplifier is therefore responsible for switching off the positive feedback loop again.

Potassium negative feedback component

A similar behaviour can be observed in the negative feedback, i.e. the potassium, component of the circuit. The source follower delays the increase of V_{mem} dependent on its bias voltage $V_{\tau K}$. Once V_{mem} becomes larger than V_{thr} in our transconductance amplifier, our output current I_K is copied by two current mirrors. The resulting current is also connected to the membrane voltage. As it pulls away from the connecting node, an increase of I_K decreases V_{mem} . You might wonder why we are using two current mirrors. First of all, we have to ensure that our current flows in the right direction to decrease the membrane voltage. Additionally, the nFET current mirror is not set to ground but to E_K . Similar to the positive feedback component, this ensure that I_K is bound by E_K .

The problem of the proposed circuit is that its behaviour is dependent on a large number of transistors. As introduced in section 3.9.2, the expected behaviour of individual transistors can vary up to 20% due to variations in their fabrication. For a large number of transistors, it becomes impossible to tune all involved variables of the circuit to get the required functionality. We therefore aim to keep our circuits as minimal as possible.

9.2 Axon-hillock circuit

How can we model the behaviour of cortical neurons, i.e. their ability to generate an action potential when the sum of the synaptic inputs reaches a certain threshold, with a minimal circuit to reduce the negative impact of device mismatch? One such circuit (which you should be able to explain at the exam) is the Axon-hillock circuit. It is an integrate-and-fire neuron model that was proposed by Carver Mead in the late 1980s. It has positive feedback (similar to sodium in neurons), negative feedback (similar to potassium), and a membrane capacitance. The circuit

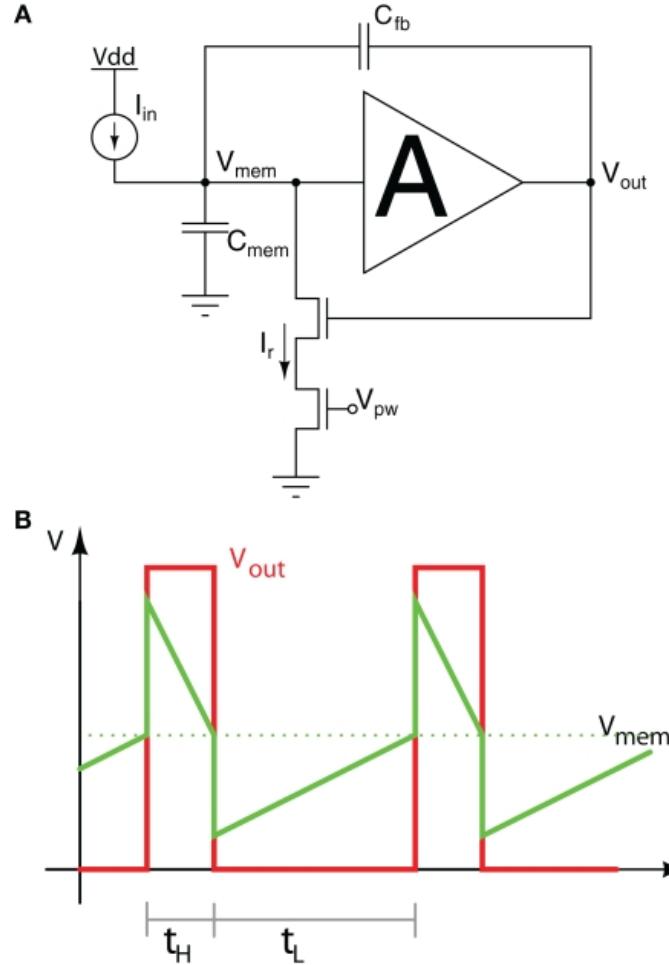


Figure 94: Axon hillock circuit.

schematic is shown in figure 94. The amplifier acts as a comparator between its input voltage V_{mem} and a threshold voltage V_{th} that is not explicitly denoted in the schematic.

The behavior of the Axon-hillock circuit can be explained as follows:

1. At the beginning, we don't have any input current and our membrane potential V_{mem} is zero. Consequently, our non-inverting amplifier also has a zero output voltage which acts as the gate voltage of the reset transistor that is turned off. For now, we ignore the positive feedback component of the circuit.
2. With constant input current, the increase in charge at the capacitor C_{mem} leads to a linear increase of V_{mem} .
3. When V_{mem} crosses the threshold voltage V_{thr} of the amplifier, the output V_{out} will go from zero to V_{dd} . We now have a positive gate voltage that turns on our reset transistor and generate a current I_r .
4. With I_r larger than I_{in} , the capacitor discharges and V_{mem} decreases again until the amplifier switches off, which turns V_{out} back to zero. Consequently, the current I_r become zero again and the charging of the membrane voltage restarts.

In reality, however, our V_{mem} is not in- or decreasing perfectly linearly but is distorted by noise. When it comes close to the threshold, the noise causes the membrane voltage to fluctuate below and above threshold and therefore leading to a fluctuating output voltage as well. How

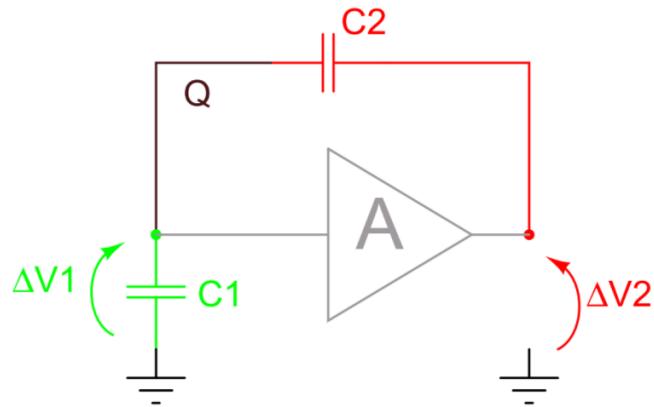


Figure 95: Capacitive divider within the Axon Hillock circuit.

can we prevent these fluctuations from happening? By introducing an additional capacitor to the circuit which acts as a positive feedback, we are able to make our membrane voltage more robust towards noise.

But how exactly does the positive feedback work? By adding a second capacitor, we get a capacitive divider circuit as shown in figure 95. Assuming that the charge within the circuit is constant, the following equation describes our circuit:

$$Q = C_1 V_1 + C_2 (V_1 - V_2) = \text{constant} \quad (201)$$

We are interested in the circuit's behaviour when our voltage changes, as is the case when our output voltage jumps to V_{dd} once we cross the threshold voltage. With $\Delta Q = 0$ for a constant Q , we get:

$$C_1 \Delta V_1 + C_2 (\Delta V_1 - \Delta V_2) = 0 \quad (202)$$

$$\Delta V_1 = \frac{C_2}{C_1 + C_2} \Delta V_2 \quad (203)$$

$$\Delta V_{mem} = \frac{C_{fb}}{C_m + C_{fb}} V_{dd} \quad (204)$$

The positive feedback inserted back into our membrane voltage corresponds to a fraction of V_{dd} as shown in (204). Thanks to our positive feedback, we push our membrane voltage away from the threshold voltage to prevent any fluctuations caused by noise.

Let's have a look at the dynamics of our circuit. We want the amplitude of our input to be linearly proportional to the frequency of our output. We denote t_L as the time between two spikes, hence $\frac{1}{t_L}$ is the spike frequency, and t_H as the pulse width of our spikes. Inbetween two spikes, our membrane voltage is below threshold while our capacitors are charging up and we only have an input current I_{in} . As our capacitors are operated in parallel, we can calculate t_L :

$$I_{in} = (C_{fb} + C_m) \frac{\Delta V_{mem}}{t_L} \implies t_L = \frac{C_{fb} + C_m}{I_{in}} \Delta V_{mem} = \frac{C_{fb}}{I_{in}} V_{dd} \quad (205)$$

We can see that the frequency $\frac{1}{t_L}$ is proportional to the input current I_{in} . We can calculate the pulse width in the same way.

$$t_H = \frac{C_{fb}}{I_r - I_{in}} V_{dd} \quad (206)$$

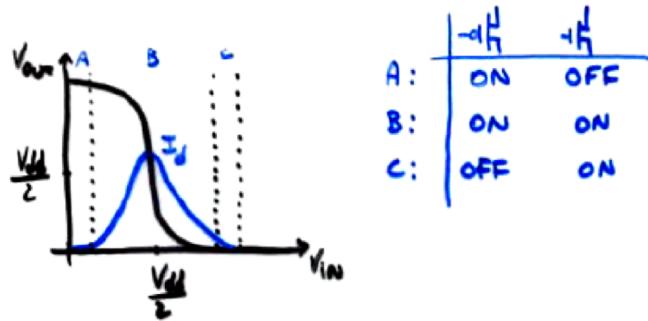


Figure 96: The slow switching time of an inverting amplifier generates a large current flow from V_{dd} to ground.

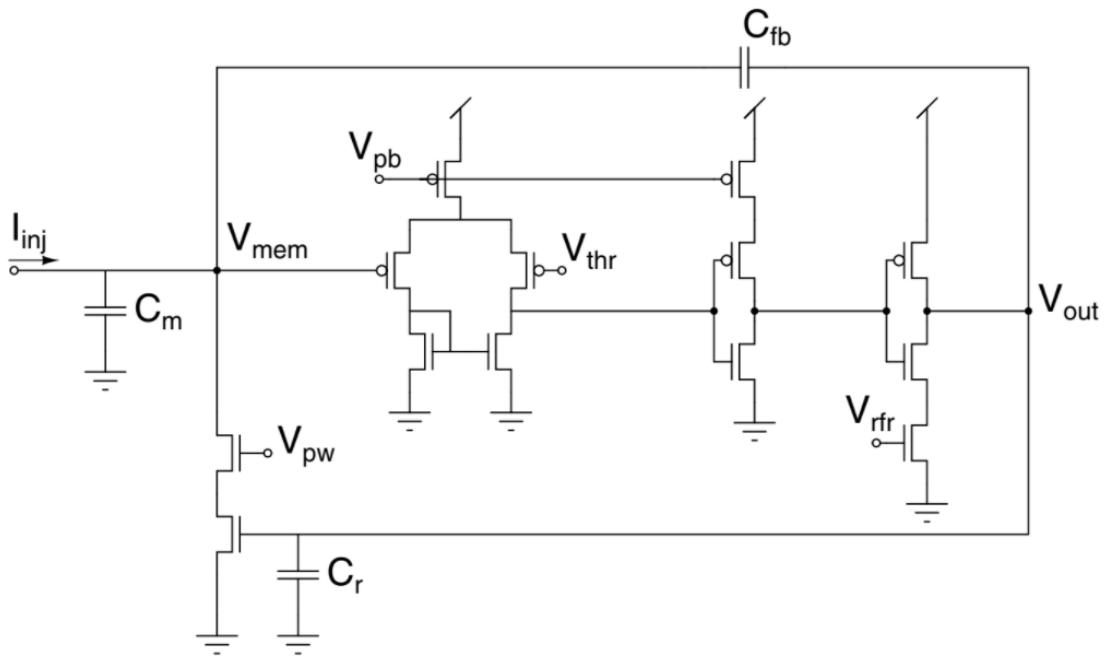


Figure 97: A more elaborate circuit of an integrate-and-fire neuron.

Assuming that the reset current I_r is significantly larger than I_{in} , we get a proportional relationship between the pulse width and $\frac{1}{I_r}$.

In the initially proposed circuit, the non-inverting amplifier was implemented with two inverting amplifiers. These, however, require several milliseconds to switch and while V_{in} is close to the threshold voltage, a large amount of current flows from V_{dd} to ground. This behaviour is visualized in 96 and it causes the Axon-Hillock circuit to have a large power consumption. In order to overcome this drawback, a non-inverting transconductance amplifier that quickly switches its output voltage without generating a large current was added in an improved version of the circuit in the late 1990s. The more elaborate circuit schematic is shown in figure 97. The two inverting amplifiers now receive the quickly switched output voltage from the transamp as an input. The resulting circuit is low-power, has an explicit voltage threshold and additionally allows to model the refractory period of real spikes. However, it still constantly burns power because of the transconductance amplifier that is always turned on.

In this chapter, we introduced silicon neuron circuits for both a conductance-based neuron

model as well as a less complex integrate-and-fire neuron model. Traditionally, these are the two main classes of neuron models. In recent years, however, there have been advances to bridge the gap between both classes and generalized integrate-and-fire neuron models have been introduced. These models are able to model many aspects of the complex behaviour of conductance-based neurons while maintain their rather simple structure. This is not relevant for the exam but an interesting outlook onto future research.

9.3 Test Yourself

You should be able to answer the following questions for the exam (mainly taken from the winter study sheet).

- What is a neuron and what are its components (synapse, soma, dendrite)?
- What types of models are used to simulate neurons?
- How does the spike-generating mechanism work?
- What is an FI curve?
- Can you draw the circuit schematic of the axon-hillock neuron?
- What are the components of the conductance-based neuron?
- What is its disadvantage?
- How can we implement a time delay?
- Why do we need a positive feedback in the axon-hillock circuit?
- What is the drawback of the circuit's initial amplifier and how does it occur?

10 Photosensors and circuits

This chapter is a sort of culmination of everything we've learnt. It is also, in my opinion the most dense and challenging to understand. Throughout writing it, I have tried my best to only give the necessary information and avoid overload of details.

Before starting, it is critical to make sure you understand well the following concepts covered previously:

- The PN Junction and Diode.

10.1 Prelude: Motivation

"Over the past 600 million years, biology has solved the problem of processing massive amounts of noisy and highly redundant information in a constantly changing environment by evolving networks of billions of highly interconnected nerve cells. It is the task of scientists—be they mathematicians, physicists, biologists, psychologists, or computer scientists—to understand the principles underlying information processing in these complex structures. At the same time, researchers in machine vision, pattern recognition, speech understanding, robotics, and other areas of artificial intelligence can profit from understanding features of existing nervous systems. Thus, a new field is emerging: the study of how computations can be carried out in extensive networks of heavily interconnected processing elements, whether networks are carbon - or silicon-based."⁵³

One of the key motivations behind the idea of Neuromorphic Engineering is to emulate the most efficient and interesting brain processes to electronics. It is highly desirable to copy the brain as it has gone through thousands of years of natural selection, it is thus likely that it would do a lot of things better than what we'd manage to engineer while overlooking it. Vision is one of the most obvious and most accessible process to emulate: the retina has been extensively studied and presents a rich myriad of layered elements that were understood to be reproducible with electronics. As mentionned in chapter 1, Misha Mahowald - in Carver Mead's lab - first brought the idea and successfully developed it. This section aims at understanding the subelements that constitute vision. We'll first look at biological vision through a brief overview of the brain and the retina. We will then proceed to study how light works and can be processed with our favourite material: silicon. Finally, we'll look at some very basic circuits that constitute building blocks of neuromorphic vision.

10.2 Light

Before getting into the specifics functioning of photosensors or the human vision, one should spend some time reviewing basic things about light itself.

Light consists of electromagnetic radiation with a dual particle-wave nature. This was pointed out by Einstein in 1905 who theorized that light is composed of discrete quanta called photons with an energy that is inversely proportional to the wavelength of the light, so that $E = hc/\lambda$. For example a photon that will be seen by our eyes as yellow has a wavelength $\lambda = 555\text{nm}$ has $E = 2.1\text{eV}$ in vacuum.

⁵³From Carver Mead's Textbook - Foreword. 1989.

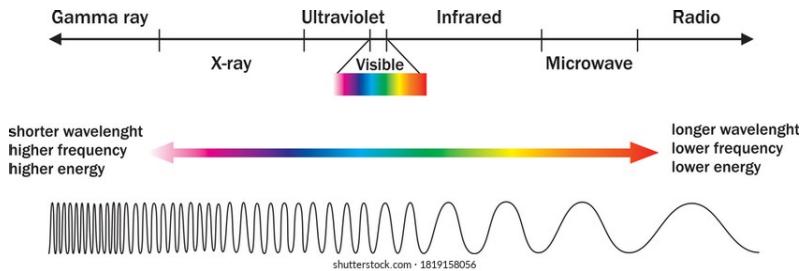


Figure 98: Range of photon wavelength. Adapted from somewhere on Google Image.

10.2.1

The photon

Photon is a basic unit of all forms of electromagnetic radiation including light. It has no mass, no electric charge, does not decay spontaneously in vacuum. In vacuum, it moves at speed of light c . We can compute the energy E and momentum P of photons:

$$E = \frac{hc}{\lambda} = h\nu \quad (207)$$

with frequency of light $\nu = c/\lambda$ in Hz.

$$p = |P| = \frac{h}{\lambda} \quad (208)$$

Its energy E and momentum P are related by $E = cp$. (p is magnitude of P). h is Planck's constant.

10.2.2 Measuring light: Radiometry and photometry

Units of light Measuring light is not done through photon energy, as it is not an informative measure of "illumination" or "brightness". Though these two are of course related to energy! The unit of visible light is the **Lumen** (noted lm). Several measures are derivatives of this unit and are more useful for everyday use. The **Lux** (noted lux) is the measure of visible light on plane surface - it is measured in lm/m^2 . The **Candela** (noted cd) is also similar to the lux but more appropriate to angular measurements: it is measured in $lm/steradian$ ⁵⁴. We typically use it to measure the illumination from light sources (e.g. lamp) as light propagates uniformly in all directions. Fun fact: prior to 2018, the basic unit of light was the Candela, but the 26th General conference on Weights and Measures redefined photon metric units. The new definition which took place on May 20th 2019, the Lumen is *defined by taking the fixed numerical value of the luminous efficacy of monochromatic radiation of frequency 540×10^{12} Hz, K_{cd} , to be 683 when expressed in the unit $lm W^{-1}$* . Understanding the units of light is not straightforward and out of the scope of these lecture notes so let's move on. 1 Lux of sunlight (because it will be different if you have a different light source as the distribution of wavelengths emitted might be different) is $\approx 4 \cdot 10^{-3} W/m^2 \approx 10^4 \text{ photons}/\mu m^2/s$.

Scene illumination and reflectance The typical illumination humans are exposed varies quite a lot: it's about 100 klux in full sunlight and 0.1 lux in moonlight. Light is reflected through surface, at different levels. Opaque materials typically absorb most of the light while the rest reflects most to all of light it receives. The average scene reflectance is about $R \approx 18\%$.

What do radiometry and photometry mean? *Radiometry* is the science of measurement of *optical radiation at any wavelength*, based simply on physical measurements. Radiant energy cannot be measured quantitatively directly, but must always be converted into some other form such as thermal, electrical, or chemical. *Photometry* is the science of measuring visible light in

⁵⁴The steradian or square radian is the SI unit of solid angle. It is used in three-dimensional geometry, and is analogous to the radian, which quantifies planar angles. A sphere has 4π steradian.

units that are weighted according to the sensitivity of the human eye. It is a quantitative science based on a statistical model of the human visual response to light - that is, our perception of light - under carefully controlled conditions.

10.3 Physics of Photosensors

The physical possibility of converting light into electrical charge is at the core of so many key processes in today's world - including biological vision itself. Before we dig into the very concepts of human vision and artificial (silicon) photosensors, it's useful to look in some detail at the basic physical phenomenon that enable such processes. Let's have a look at these:

10.3.1 Photoelectric Effect

The photoelectric effect⁵⁵ is the fundamental physical phenomenon underlying pretty much everything we can do with light: from vision and cameras to the photo-voltaic energy generation. Briefly, it is the emission of electrons when electromagnetic radiation, such as light, hits a material. Electrons emitted in this manner are called photoelectrons and are very useful to say the least.

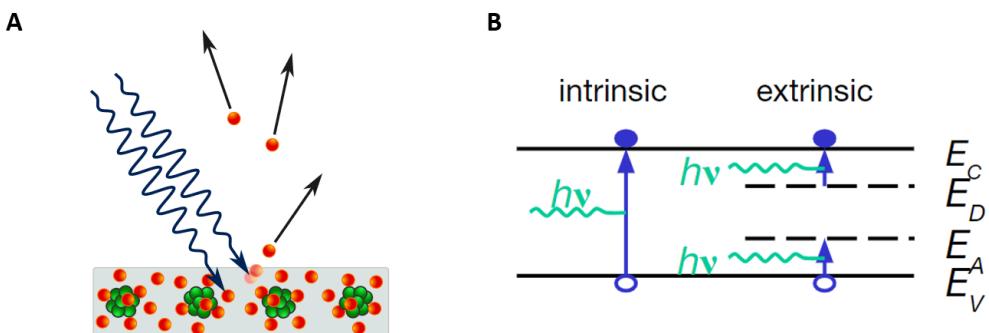


Figure 99: A) Basic principle of Photoelectric effect. Incident photons represented with the zigzag blue arrows and emitted photoelectrons with the black straight arrows. (Adapted from somewhere on Google Image). B) Photoelectric effect with Band Energy Diagrams applied to Semiconductors. We can see that energy needed for photoelectron emission is higher for intrinsic semiconductor than extrinsic semiconductor. E_D and E_A are Fermi level of N-Type and P-Type semiconductors, respectively. (Adapted from Lecture Notes).

Figure 99.A) shows the basic intuition behind the photoelectric effect. Incident photons provide some energy that will be enough to take electrons in Valence band out of the valence band and into the conduction band, thus becoming "free" to move! This is best shown in Figure 99.B). As we established before, the energy of a photon (in vacuum) can be calculated from its wavelength $E_{\text{photon}} = h\nu$.

We can think of the incident light as a stream of photons with a given energy determined by the light frequency. When a photon hits a metal surface, the photon's energy can be *absorbed* by an electron in the metal. This is the same energy principle that we discussed in Chapter 2 when talking about conduction and valence shells, where some energy needed to be brought to free electrons. And exactly as we discussed in chapter 2, there is a minimum energy (so minimum frequency)

10.3.3 Optical Absorption

Now that we established the basic principles behind photoelectron emission, which comes from absorption of photon energy, we should look at some parameters that influence the absorption of photons within a certain material.

⁵⁵Most of this explanation is based on the excellent presentation of the topic on Khan Academy: <https://www.khanacademy.org/science/physics/quantum-physics/photons/a/photoelectric-effect>.

When photons are incident onto a surface, they have a certain probability of generating an electron-hole pair, in any "slice" of Silicon. A slice is simply a layer of silicon atoms, that we can also define as an increment depth quantity δx . As a consequence of this non 0 probability of generating an electron hole pair, the number of photons decreases exponentially within the depth of the surface - the more you go inside the less photos you find because a lot of them have already been absorbed. Every slice δx removes some fraction. (See figure 100.A). This actually varies a lot with wavelength (and thus energy). The longer the wavelength, the lower the energy and the further photons can penetrate before electrons hole pairs are made. This may seem counter-intuitive but it makes sense: it's less likely that low energy (high wavelength) photons create electron hole pairs than higher energy ones - so photons manage to go further inside. This is illustrated in Figure 100.B).

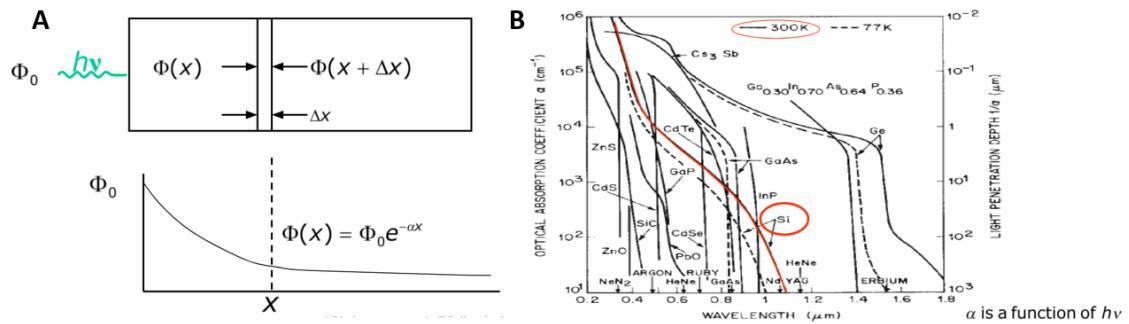


Figure 100: A) Schematic of varying photon flux as a function of depth. Top image shows the principle of "Slice" and bottom shows the exponential decay of photons as a function of material depth. B) Absorption as function of wavelength in different material, with Silicon highlighted in red. We can clearly see that optical absorption decreases with increasing wavelength (and thus lower energy). Adapted from Lecture Notes.

Here are the maths underlying the concept. We define photon flux as a function depth x as $\phi(x)$. The number of photons absorbed within δx is given by the absorption coefficient *alpha*:

$$\frac{d\phi(x)}{dx} = -\alpha\phi(x) \quad (209)$$

This differential equation can easily be solved:

$$\phi(x) = \phi_0(x)e^{-\alpha x} \quad (210)$$

The ϕ_0 property is important and factor of a few things, such as Reflection R , cross section A and incident optical power P_{opt} . I mention these only for reference, knowing this is not required for the exam - though this should intuitively make sense.

$$\phi_0 = \frac{1-R}{A} \cdot \frac{\lambda}{hc} \cdot P_{opt} \quad (211)$$

In silicon, the longest wave length that can create photoelectrons is $1.1 \mu m$ - just above visible domain. Huh, once again, we're in luck, Silicon happens to have just what we need.

10.3.4 Quantum Efficiency

One last thing to talk about before we start looking at the details of some actual sensors: quantum efficiency. Quantum efficiency, η (or *QE*) is defined as the number of electron-hole pairs generated for each incident photon (always ≤ 1 !!). This is particularly important to evaluate the quality of photosensors, and also is a very important measure for solar energy things. Of course you want a high number of electrons (and thus current) to be emitted from light. This is really what needs to be optimized in a photosensor. This concept should become clearer when talking about specific photosensors, so hold it if it's not clear yet.

$$\eta = \frac{I_{ph}}{q} \cdot \frac{hc/\lambda}{P_{opt}} \quad (212)$$

With I_{ph} photogenerated current, q number of carriers (electron or holes) and P_{opt} incident optical power.

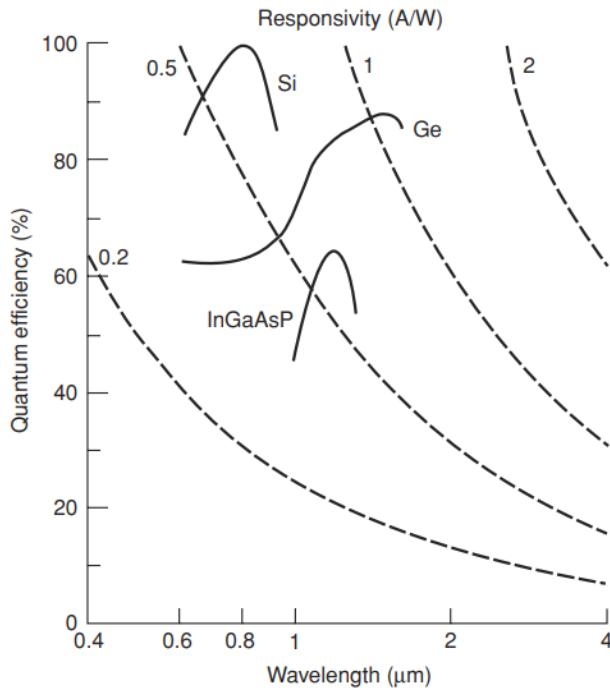


Figure 101: Quantum efficiencies and responsivities of photosensors fabricated from different semiconductors. Silicon exhibits a very good quantum efficiency peaking in the near infrared which may approach 100% in a certain spectral range. Adapted from Textbook.

10.4 Interlude: Human Vision and the Retina

Before jumping into silicon photoreceptors, let's have a very brief look at human vision. We'll mostly focus on understanding our "photoreceptor" which is the retina and won't go into the processing details of the visual cortex which, despite having been extensively documented since the work of Hubel and Wiesel in the 1960s⁵⁶ is still an active domain of research due to its complexity.

10.4.1 General architecture of the visual system

Some things you should know about the architecture of the visual system:

- Retinal image is inverted and reversed compared to the visual field
- The axons of retinal ganglion cells form the optic nerves.
- At the *optic chiasm*, axons from the temporal halves of each retina continue into the optic tract of the same side. Axons from the nasal halves cross to the optic tracts on the opposite side. **So the myth that vision is systematically processed on the opposite side of the corresponding eye is false.** But the visual field is separated in two on each eye.

⁵⁶Hubel and Wiesel worked together at Harvard on visual processing in the cat. Their research, spanning over two or three decades, and for which they both received a Nobel Prize, remains some of the most influential works in Neuroscience to date.

- Most of the primary visual cortex (also called V1) is on the medial surface of the human brain (see figure). This is one of the key place where visual input is processed in the brain.

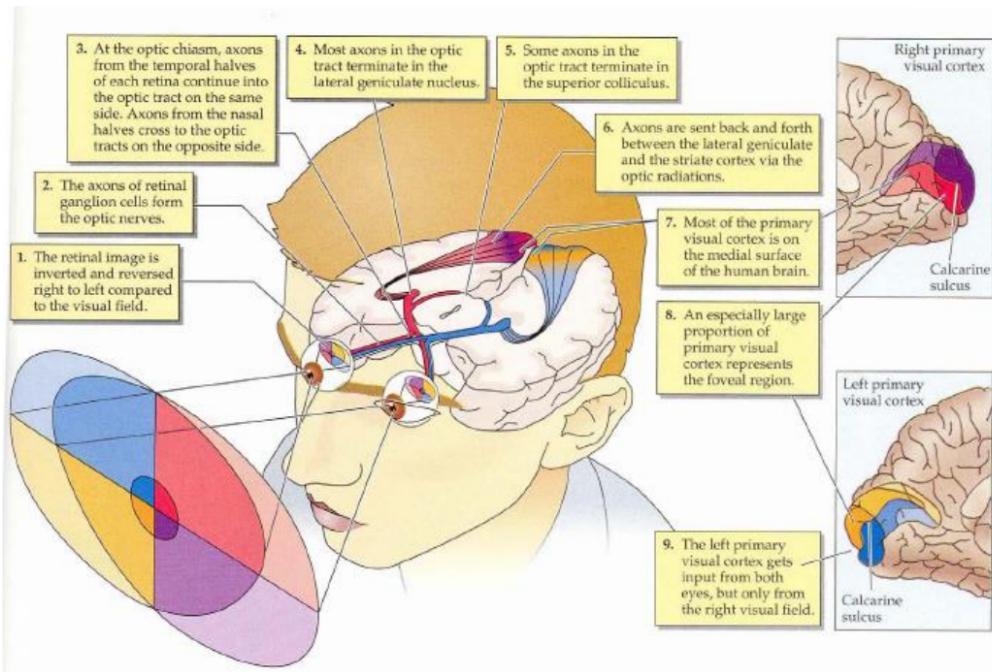


Figure 102: Simplified structure of the Visual system. Adapted from somewhere on Google Image.

10.4.2 The human eye

Not much to say besides what's shown on figure 103, this is just about organizational things. The important bit is where the retina is, where interesting things happen.

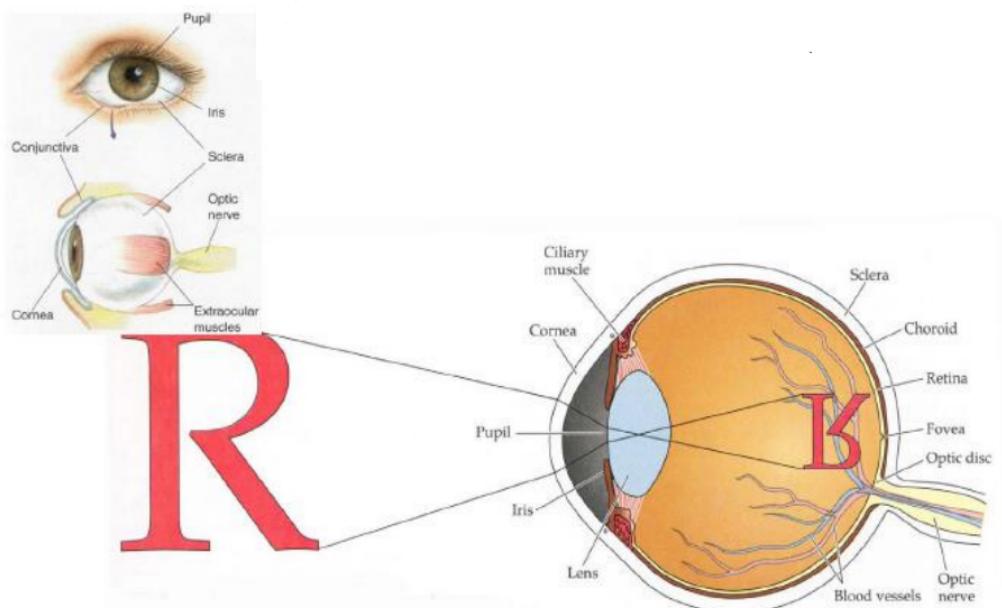


Figure 103: Simplified structure of the eye. Adapted from somewhere on Google Image.

10.4.3 The retina

The retina is where the photoreceptors are located. The important processing happens first through the cones and rods - the eye's photoreceptors, followed by the bipolar cells and then ganglion cells, which send the signal to the brain. This may be a bit counterintuitive when first looking at figure 104: light **arrives first in the photoreceptors**, which are at the *back* of the retina. The photoreceptors will then transmit to the ganglion cells through the bipolar cells. Finally, the ganglion cells carry information from the retina to the brain. If you pay attention to the "wiring" in the figure, this should become clearer.

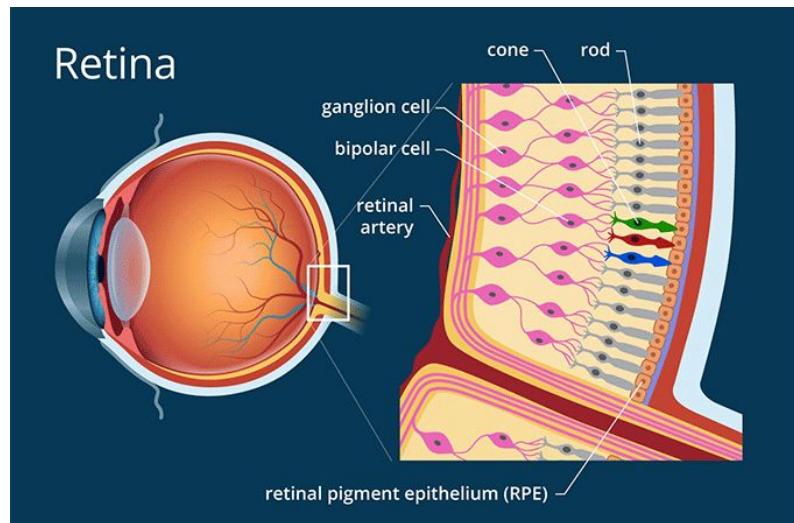


Figure 104: Simplified structure of cells in the Retina. Light reaches first the photoreceptors (cones and rods), and transmit, through bipolar cells, to the ganglion cells which then transmit to the brain. Adapted from somewhere on Google Image.

10.4.4 Human photoreceptors: Rods and Cones

The most interesting part for NE1 is understanding the photoreceptors, their structure and organisation. There are two main types of photoreceptors in the retina: Rods and Cones. They are very particular types of neurons. Rods (≈ 120 Millions) are very sensitive and are most useful in diminished light. They "saturate" at high illumination. Cones on the other end are much less numerous (≈ 5 Millions) and are only sensitive to brighter light. There are also three different types of cones, each being specifically receptive to a range of wavelength (colors). A summary of the distribution and optimal wavelength for each type of photoreceptor is shown in figure 105.

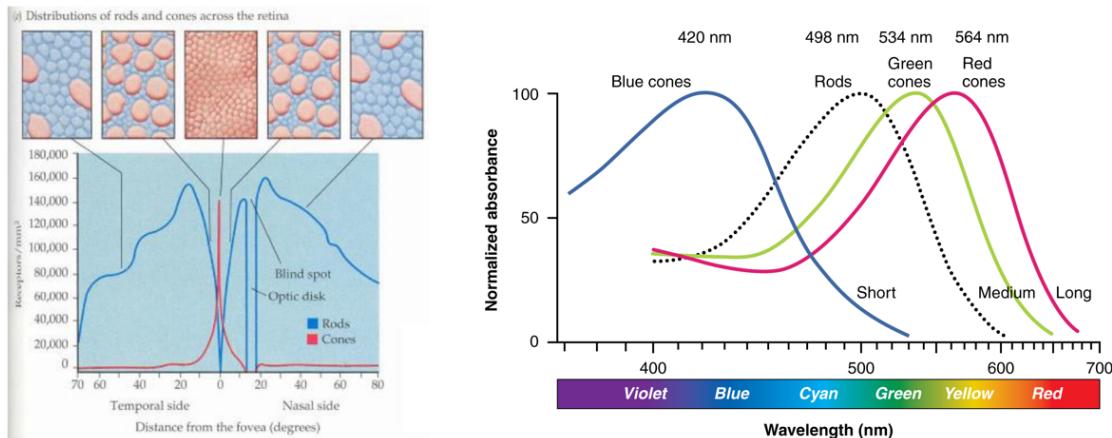


Figure 105: A) Distribution of Cones and Rods in the retina. B) Wavelength absorbance of different types of photoreceptors of the retina. Adapted from somewhere on Google Image.

We should also note that there are no rods but many cones in the *fovea*: it is the region in our retina that provides the highest acuity of vision (see Cones peak region in figure 105). There are also no photoreceptors on the optic nerve (see figure 103) which constitutes a *blind spot*.

10.4.5 Photo receptor structure and function

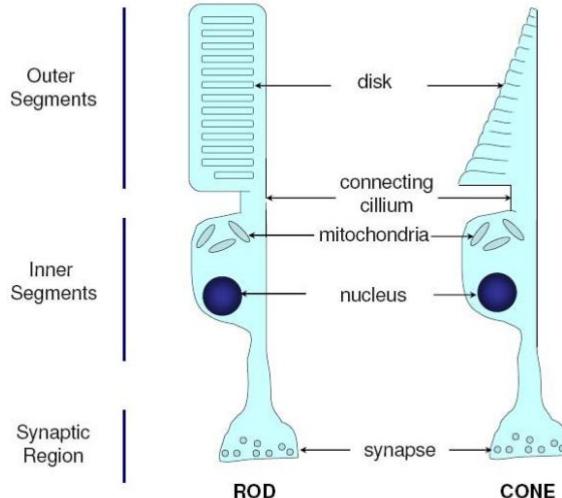


Figure 106: Structure of Rods and Cones. Adapted from somewhere on Google Image.

As can be seen in figure 106, rods and cones both possess "outer segment disks" which are filled with *opsin*, a protein which absorbs photons as well as voltage gated Sodium (Na) channels. These proteins are actually slightly different in cones and rods: cones carry *photopsin* whilst rods carry *rhodopsin*. However, both actually function very similarly: photons being in contact with the opsin proteins are absorbed by the latter: this is due to the photoelectric effect!. This generates a biochemical cascade (including a decrease in *Glutamate* release) leading to a hyperpolarization of the photoreceptor cell and eventually to triggering an *action potential*. Bipolar cells subsequently respond to this hyperpolarization and things go on from there to the ganglion cells etc... Figure 107 shows the change in current after light trigger (at time 0) through the photoreceptors.

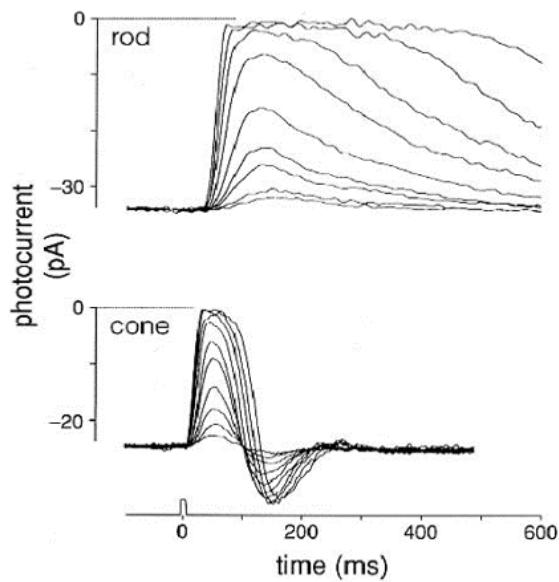


Figure 107: Cones and Rods response to light. Adapted from somewhere on Google Image.

Let's give a summary of the key points about photoreceptors, because this is what should be remembered:

- Photoreceptor cells are depolarized in the dark
- Light hyperpolarizes the cell, with activates the next cell in the neural pathway: bipolar cells. This is through the photoelectric effect.
- In the dark, levels of Glutamate are high and allow a *steady* inward *dark* current. This keeps the cell at a depolarization value of about -40 mV.

10.5 Silicon Photosensors

10.5.1 Common principles of Photosensors

The growth in the market for optical communication and electronic imaging has brought a huge importance to the photosensor as optoelectronic interfaces. Consequently, a wide range of photosensors designed for different applications are available. Here, we will consider only photosensors that can be fabricated with semiconductor processes used for the implementation of transistor-based electronic devices. Artificial Photosensors work just like the biological ones we've seen in the retina: they convert electromagnetic radiation (photons) into a different physical form using photoelectric effect.

In a semiconductor, an incident photon and therefore its energy can be absorbed by an electron; a process known as the inner photo-electric effect. A photon with an energy larger than or approximately equal to the bandgap energy can excite an electron from the valence band into the conduction band. In the energy-band diagram, this process corresponds to the generation of an electron-hole pair⁵⁷. Illumination of a semiconductor therefore increases the concentration of mobile charge carriers above the thermal equilibrium value in the exposed area. If the motion of the carriers is driven by diffusion only, then the generation is balanced by recombination. However, in the presence of an electric field, which typically leads to drift current, electrons and holes can separate and some of the separated carriers contribute to an electrical output signal. This phenomenon is called photoconduction. The simplest device that implements the phenomenon is the photoconductor, which is a slab of semiconductor in an externally applied electric field. Photoconductors exhibit a large **dark current**, which is a background current

⁵⁷The notation of "pair" might be a little confusing. It means that the electron now has enough energy to move around freely and when it does a hole remains in the shell. We know have a pair of a free electron and its hole.

which is still present in the absence of optical stimulation. This current is due to the relatively large doping concentration used in most modern semiconductor processes. This doping results in high conductivity values and a poor signal-to-noise ratio of photoconductors. There are also other types of photoreceptors such as the photogate or the phototransistor but we're not looking at these in NE1. Let's just have a look at the photoconductor though before getting to the important part of the photodiode.

10.5.2 Photoconductor

This is not something we focus on in NE1, but it's the simplest form of photoreceptor. Let's briefly have a look.

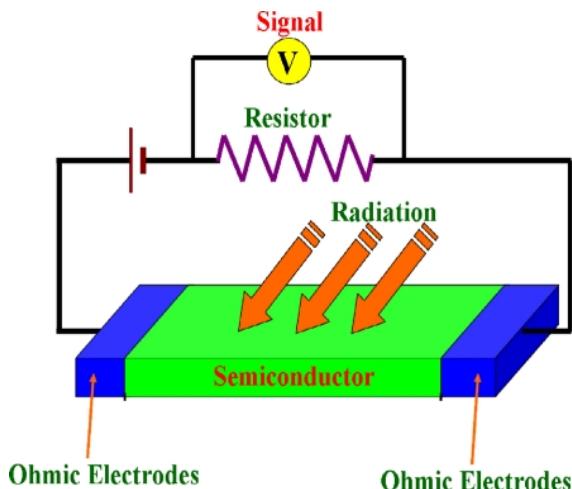


Figure 108: Photoconductor inside a circuit. Adapted from somewhere on Google Image.

It's simply a piece of semi conductor that receives light and conducts current as a consequence. The conduction rate will be proportional to multiple things, including the cross sectional area of the piece of semi conductor, it's depth, the quantum efficiency, the incident optical power, and of course the wavelength and energy of incident light. Charge is then transported through the externally applied electric field (drift current).

10.5.3 Photodiode

The photodiode ⁵⁸is the most important photosensor studied in NE1. This is the one we should know about the most.

A diode is a much more suitable photosensor, because it has a depletion region with a low conductivity and a built-in electric field. The presence of a depletion region substantially reduces the dark current, while the built-in electric field in the depletion region performs charge separation even in the absence of an externally applied voltage. This generates, as seen in figure 109 a reverse current (this is not the same reverse bias which refers to voltage you apply at the nodes). We say that the current is reverse because electrons flow from p-region to n-region (and hence current from n-region to p-region), which is the opposite of normal current and electron flow direction.

⁵⁸I highly recommend watching Khan Academy's explanation on the topic, which does a brilliant job at explaining the reverse bias aspect of generated photocurrent. <https://www.youtube.com/watch?v=KgKcbW77txY>

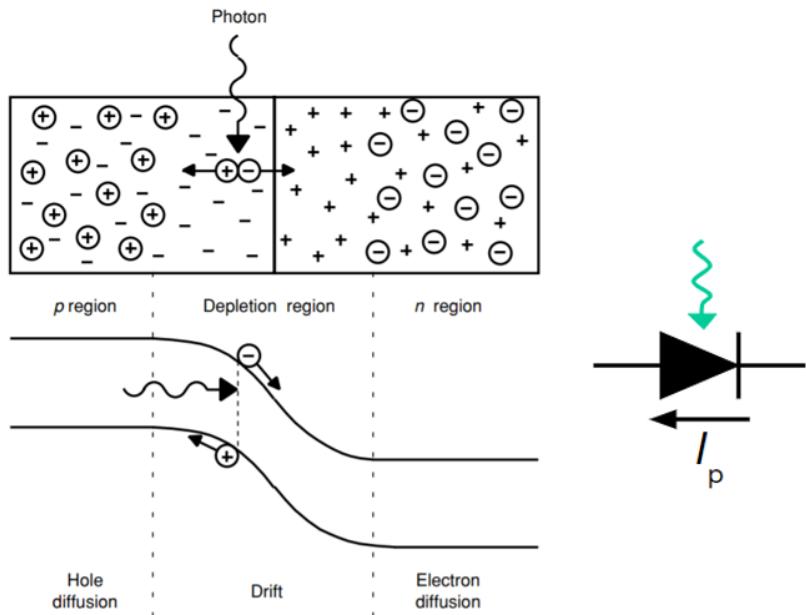


Figure 109: Principle of operation of a photodiode. Electron-hole pairs generated by incident photons in or within a diffusion length outside the depletion region become separated and contribute to a reverse generation current. Adapted from textbook.

Why do we typically operate photodiodes in reverse bias? Remember from chapter 2 that the depletion region is much larger when operated in reverse bias compared to forward bias, and the leakage current is also significantly smaller. When operating photodiodes, we want to have something very sensitive to incident light, so we want the largest possible surface area (in this case the depletion region) to receive incident light, and the resulting photocurrent not to be drowned into already existing current. This is why operating it in reverse bias makes sense.

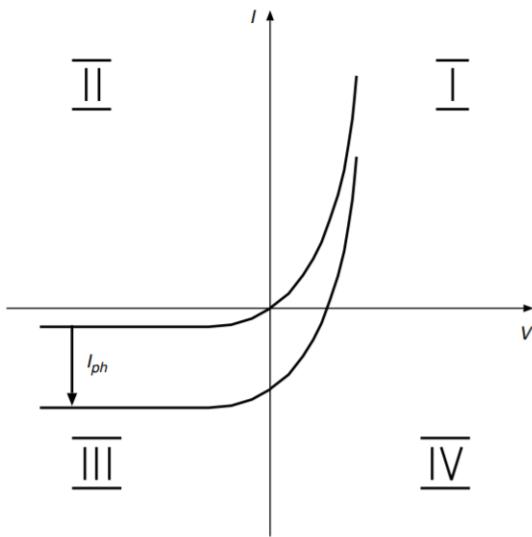


Figure 110: Steady-state current-voltage characteristics of a photodiode. The upper curve is the normal diode characteristic (dark characteristic). The lower curve shows the characteristic under illumination. Photodiodes are usually operated either in quadrant III as photosensors, or in the quadrant IV as solar cells. Remember from chapter 2 that reverse bias is quadrant II and III ($-V$), and forward bias is I and IV ($+V$). Forward current is I and II and reverse current is III and IV. Adapted from textbook.

A photodiode has two principal modes of operation, depending on its application. If the photodiode is used to convert optical power into electrical power it is called a solar cell and operated in the photovoltaic mode. In this mode, a load is connected between the two terminals, such that a reverse current flows through the diode in the presence of a forward voltage. A solar cell is thus operated in the quadrant IV of the current-voltage characteristic. The generated power is given by the product of the reverse current and the forward voltage. Commercial solar cells are complicated devices, which are optimized with respect to optical-to-electrical power-conversion efficiency that is typically between 10% and 20%⁵⁹.

In the other mode of operation, the photosensing mode, the photodiode is used to estimate the photon flux. In steady state, the diode is typically either open-circuited and the forward voltage is read out or a reverse (or zero) bias is applied to the diode and the reverse current is read out. In the latter case, the photodiode is operated in the quadrant III of the current-voltage characteristic. Here, the photodiode is quite a good current source, because the photocurrent is almost independent of the applied reverse bias.

The important thing to remember in terms of dynamics (because no need to get into the maths) is that we obtain by operating a photodiode in such a way a **current that is linearly proportional to light intensity**.

Considerations about photodiodes Photodiodes designed to be operated in the continuous-current photosensing mode are usually optimized with respect to their quantum efficiency and their response time, which are two partly conflicting requirements. The quantum efficiency can be optimized by applying an anti-reflection coating to the semiconductor surface to reduce R and by generating a thick depletion region close to the surface. The response time can be kept small by minimizing the junction capacitance, the carrier transit time through the depletion region, and the carrier diffusion time to the depletion region. A small junction capacitance requires a thick depletion region, while a short transit time favors a thin depletion region and a large drift velocity. Diffusion times to the depletion region can be minimized if the depletion region extends close to the surface. It is thus advantageous to operate a photodiode at a large reverse bias in order to increase the depletion region width and the drift velocity. The depletion region width can also be

⁵⁹High-efficiency solar cells are built with a layering of different semiconductors of different bandgap

increased by having a low impurity-doping concentration in the junction region, as is done in most commercially-available photodiodes. Such photodiodes are known as p-i-n photodiodes, because they have a (nearly) intrinsic region between the n-type and p-type region. In the photodiode operation range described so far, the quantum efficiency is smaller than unity: Each photon cannot produce more than one electron-hole pair. However, if a diode is operated in the avalanche multiplication regime in the vicinity of reverse junction breakdown, the photogenerated carriers multiply in the depletion region due to impact ionization and the quantum efficiency can be significantly larger than unity. Avalanche photodiodes are photodiodes designed to be operated in this domain. They have small response times and better signal-to-noise ratios than normal photodiodes. A normal semiconductor process provides very poor avalanche photodiodes with instability and matching problems. A photosensor with an internal gain mechanism that may be implemented more efficiently with standard processes is the phototransistor, described in the following section.

10.6 Operations of Photoreceptors

10.6.1 Dark Current

Photodiodes leak even in darkness. Here are some key points to know about Dark Current:

- Typical process leaks about 1 nA/cm^2 at 25 degrees Celsius.
- This actually corresponds to moonlight scene illumination. So that means it's difficult to differentiate between moonlight and complete darkness!
- Most of the leakage comes from the edges of the junctions, where leakage is 10-100x higher. This makes sense considering the absorption principle discussed previously.
- Leakage current doubles every 6 to 8 degrees. Don't use it at night in summer!

10.6.2 How to estimate incident light on the chip

So this is to reach estimates of current from given light intensity. Here are some considerations before reaching the final equation:

- $1 \text{ Lux} \approx 10^4 \text{ photons}/\mu\text{m}^2/\text{s}$
- Moonlight 0.1 Lux; Office light 500 lux; Full sun 10^5 lux.
- Average scene reflectance $\approx 18\%$
- You operate it using a lens (with focal length f). This allows you to concentrate the incident photons into the smaller area your photodiode is located in.
- Lux falling on chip is $\frac{1}{4f^2}$ imaged from white surface. Typical fast lenses⁶⁰ used with have $f = 1.4$.
- Quantum efficiency of photodiodes is about 0.5.

We can now establish an equation that works for estimating amount of light (in Lux) falling on the chip I_{chip} (in Lux) as a function of scene illumination $I_{scene}[\text{Lux}]$ (in Lux):

$$I_{chip}[\text{Lux}] = \frac{I_{scene}[\text{Lux}] \cdot R \cdot QE}{4f^2} \quad (213)$$

This typically yields that it's only a small fraction of scene illumination (1/30th) that falls on the chip.

⁶⁰https://en.wikipedia.org/wiki/Lens_speed

10.6.3 Why is a log response desirable?

During the lab, we studied photoreceptor with logarithmic response to light intensity. A logarithmic response has the following advantage:

- Static scene illuminance I appears as additive term in the output, formed from the product of I and scene reflectance R : $\log(RI) = \log(R) + \log(I)$
- Differences between photoreceptors over space or time leave only the reflectance variations: $\Delta\log(IR) = \Delta\log(R)$
- Reflectance variations are object properties, which are useful for vision.
- The log is also very compressive, allowing wide dynamic range within a power supply rail - as long as mismatch can be tolerated.

10.7 Logarithmic Photoreceptor

There exists several types of logarithmic photoreceptor circuits that exhibit different properties, these are shown in figure 111.

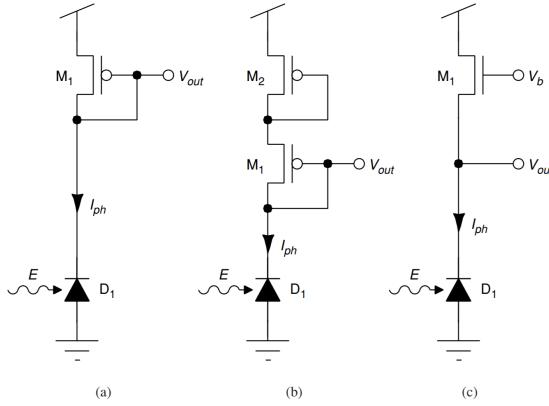


Figure 111: Photosensors with logarithmic irradiance-to-voltage conversion for subthreshold photocurrents, consisting of a photodiode and a current-to-voltage conversion stage implemented as (a) a diodeconnected MOSFET, (b) two diode-connected MOSFETs in series, (c) a MOSFET in sourcefollower configuration. Adapted from Textbook.

I include the transfer functions for these circuits here without derivation. This is typically the kind of things that I believe they may ask us to derive in the exam, though I leave it as an exercise for the reader :).

$$\text{Diode Connected : } V_{out} = V_{dd} - \frac{U_T}{\kappa} \log\left(\frac{I_{ph}}{I_0}\right) \quad (214)$$

$$\text{Double Diode Connected : } V_{out} = V_{da} - U_T \frac{\kappa + 1}{\kappa^2} \log\left(\frac{I_{ph}}{I_0}\right) \quad (215)$$

$$\text{Source Follower : } V_{out} = \kappa V_g - U_T \log\left(\frac{I_{ph}}{I_0}\right) \quad (216)$$

UNCLEAR: I_{ph} is the "static scene illuminance" (in Lux).

10.8 Reasoning through Source Follower Logarithmic Photoreceptor

Now let's look in details at the circuit that we have studied in class and that we most likely will be asked about: the source follower photoreceptor

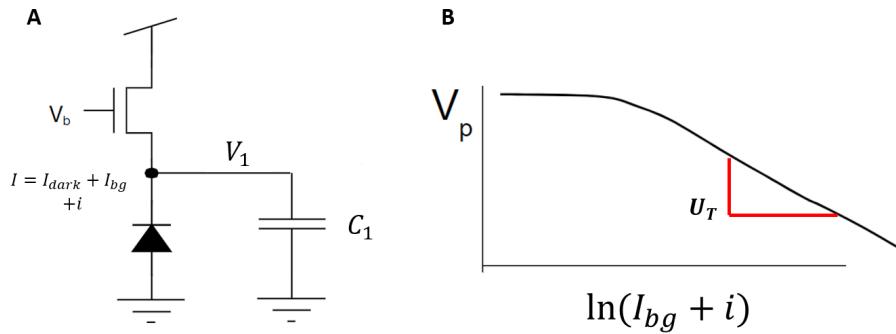


Figure 112: Source Follower Circuit with parasitic capacitance considered. I_{dark} is the constant dark current, I_{bg} the static scene luminance i signal luminance (varying part). Resulting current $I = I_{dark} + I_{bg} + i = I_0 e^{\frac{\kappa V_b - V_1}{U_T}}$. Adapted from Lecture notes.

The difference between the circuit shown in figure 112 and figure 111.C is the capacitor. In reality, the capacitance is also present in figure 112.C (A and B as well for the matter) but it is omitted from the schematic. The capacitance that you see is the "parasitic capacitance" which we have already encountered before. It is indeed present in all electrical components and need sometimes be considered. This is the case here as we need to consider the speed at which response is obtained, which is an obviously important criteria for photosensors.

Let's first analyze the source follower circuit. Tobi's explanation on the topic is extremely clear, so I'll just copy (almost) word to word what he said in the lecture. It's best to proceed step by step, so let's do it like this:

1. When light shines on the photodiode you see, a current is generated. This current is linearly proportional to the incident light.
2. If there is no light, you have some dark current.
3. When light shines, because of conventional current source, the photodiode acts as a constant current *sink*.
4. When light shines, the photodiode sinks currents, and in way, *forces* the current to sink. That means that the top transistor is forced to adapt itself to match the current from the photodiode. As we assume that the gate voltage is unchanged, the source voltage of our top transistor (V_1/V_p) must change to match the photodiode. The more light shines, the higher the current, and hence the lower V_1/V_p need to be (to reach a higher V_{gs}). We can thus write the following equations, by taking I as the current flowing through the photodiode, equal to the dark current + the photocurrent:

$$I_0 e^{\frac{\kappa V_b - V_1}{U_T}} = I \quad (217)$$

5. Since the current flowing in in the transistor is *exponential* to the V_{gs} , then it follows the the source voltage changes in *logarithmic* fashion with the changing current. This yields@

$$V_1 = \kappa V_b - U_T \ln\left(\frac{I}{I_0}\right) \quad (218)$$

We can also derive from this the graph which is seen in 112, the slope of the linear part is U_T

6. Every photodiode has an associated parasitic capacitance associated to it - this also happens to low pass filter the signal here (because there is a conductance at the node and then the capacitance, so it's an RC configuration).

7. There is thus a time constant associated with the node, which is related to the capacitance and the conductance of the node. The conductance of the node is "how much does the current in the node change when changing the voltage".
8. To evaluate the conductance, let's remember that the photodiode behaves like a constant current sink (assuming constant light), so this current does not change with the voltage (V_p). However, the current in the transistor is heavily (exponentially) controlled by the voltage (V_p).
9. The source conductance is therefore simply the "source conductance" of the subthreshold transistor, which we've derived before. This is $g_s = I/U_T$.
10. We thus now have the time constant of our circuit: $\tau = RC = C/G = \frac{C \cdot U_T}{I}$.
11. The time constant is inversely proportional to the photocurrent. In other words, **the brighter the light, the faster the circuit behaves**. This is a problem because it means our circuit will not behave properly in low light conditions. To deal with this problem, we typically go for a more practical transimpedance log photoreceptor.

10.8.1 Time domain response of source follower response

1. The current flowing through the capacitor must be the current sunk by the photodiode minus the current flowing through the transistor. Let's omit U_T for simplicity.

$$Cv_1^{61} = e^{\frac{\kappa V_b - V_1}{U_T}} - I \quad (219)$$

2. Small signal analysis yields:

$$Cv_1 = -g_s v_1 - i \quad (220)$$

3. Knowing that $\tau = c/g_s$, we have the transfer function:

$$(\tau s + 1)v_1 = -\frac{i}{g_s} \quad (221)$$

If we compute the magnitude of the transfer function:

$$\left| \frac{v_1}{i/g_s} \right| = \left| \frac{1}{\tau s + 1} \right| \quad (222)$$

This is the transfer function of a **lowpass filter!** Note that the input is current and we obtain volts, which is why we speak of transconductance g_s .

10.9 Reasoning through Transimpedance Logarithmic Photoreceptor

This is common exam question!

⁶¹Dot means derivative - here with respect to time

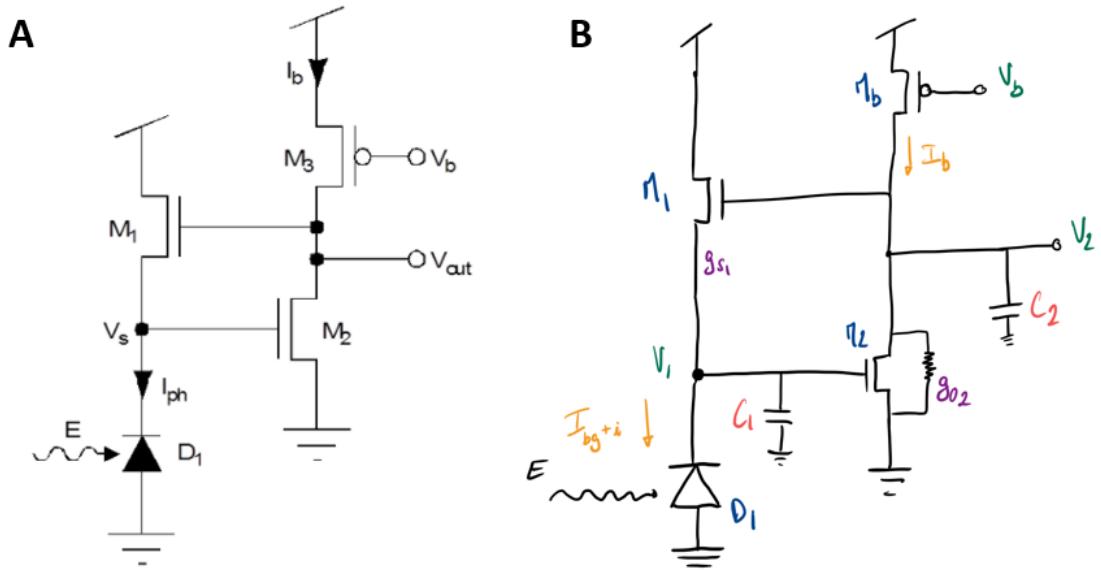


Figure 113: A) Basic Transimpedance Logarithmic Photoreceptor diagram. We only need this for DC response. B) Transimpedance Logarithmic Photoreceptor which considers the parasitic capacitance and conductance, which need to be considered for the time frequency response. We will be using naming conventions of B) in our analysis. Adapted from Lecture notes.

1. Circuit follows the same logic, except that instead of M_1 having a fixed gate voltage as in the source follower, we have some kind of inverting amplifier setting that controls it.
2. Photodiode still acts as a constant current sink as in the previous source follower arrangement.
3. M_3 is a saturated pFET which draws a constant current I_b .
4. Similarly to before, V_1 , gate voltage of M_2 , must be at a value that matches the current I_b flowing in the right branch. It follows that V_1 must match V_b . That is, $V_{dd} - V_b = V_1 - 0$.
5. V_2 , gate voltage of M_1 must provide a V_{gs} to M_1 that will match $I_{bg} + i$. Note that V_{gs} of M_b is $V_2 - V_1$
6. **Important:** A popular exam question to take some values for a starting voltage point and reason you way through the rest. This is particularly true where we can reason through the DC output of this circuit:
 - (a) Let's for example assume that $V_1 = 0.7V$
 - (b) We thus from 1. need to have $V_{dd} - V_b = 0.7V$.
 - (c) Let's assume that we have a saturated transistor instead of a photodiode. This allows us to have a voltage to reason with rather than current, though both effectively do the same thing: provide constant current from constant input. This is just for the sake of the analysis because we can now choose a gate voltage for this transistor M_0 that we call V_{g0} .
 - (d) If we take gate voltage of this transistor to be $V_{g0} = 0.4V$, then it follows from 2. that $V_2 - V_1 = 0.4V$.
 - (e) We thus have value of $V_2 = 1.1V$!
7. The DC response of this circuit is:

$$V_2 = k^{-1}(V_1 + U_t \log(\frac{I_p h}{I_0})) \quad (223)$$

Note that this doesn't consider the parasitic capacitance and conductance as we're doing the DC output.

8. **Summary:** This circuit is an improvement of the previous one, where we enhance the response speed by adding a high-gain negative feedback loop from the source to the gate of the MOSFET in the previous source follower configuration. The voltage output signal then appears at the gate of the MOSFET, while the source and therefore the voltage across the photodiode is practically clamped.

10.9.1 Time domain response of transimpedance logarithmic Photoreceptor

1. From the circuit we can get the following differential equations for V_1 and V_2 :

$$C_1 \dot{V}_1 = I_0 e^{\frac{\kappa V_2 - V_1}{U_T}} - I_{ph} \quad (224)$$

$$C_2 \dot{V}_2 = I_b - g_{o2} V_2 - I_0 e^{\frac{\kappa V_1}{U_T}} \quad (225)$$

2. In the small signal regime, we can simplify this to:

$$C_1 \dot{v}_1 = g_{m1} v_2 - g_{s1} v_1 - i \quad (226)$$

$$C_2 \dot{v}_2 = -g_{o2} v_2 - g_{m2} v_1 \quad (227)$$

with $g_{s1} = \frac{-I_{ph}}{U_T}$, $g_{m1} = \frac{\kappa I_{ph}}{U_T}$, $g_{o2} = \frac{I_b}{V_e}$, $g_{m2} = \frac{\kappa I_b}{U_T}$

Note that we also neglected I_b here, as it is constant and we are interested in the small changes only.

3. By defining the time constants $\tau_1 = \frac{C_1}{g_{s1}}$ and $\tau_2 = \frac{C_2}{g_{o2}}$ and remembering how to use Laplace transform to solve linear differential equations (just replace the derivative with multiplication by s), we can rewrite this to:

$$(\tau_1 s + 1) v_1 = \frac{g_{m1}}{g_{s1}} v_2 - \frac{i}{g_{s1}} \quad (228)$$

$$(\tau_2 s + 1) v_2 = -\frac{g_{m2}}{g_{o2}} v_1 = -A v_1 \quad (229)$$

where A is the gain of the output transistor by definition.

4. Solving this system of polynomial equations, we get:

$$v_2 = \frac{-\frac{A}{g_{s1}}}{\tau_1 \tau_2 s^2 + (\tau_1 + \tau_2)s + 1 + kA} \quad (230)$$

10.10 Laboratory : Photoreceptors

In this lab we will compare the source-follower (SF) photoreceptor with the unity-gain active transimpedance feedback (TI) photoreceptor.

10.10.1 Static DC responses

First, we study the response of these two circuits to a static input photocurrent. These two circuits have symmetrical responses at the output voltage. To put it simply, with less light the SF photoreceptor outputs a bigger voltage and viceversa for the TI photoreceptor.

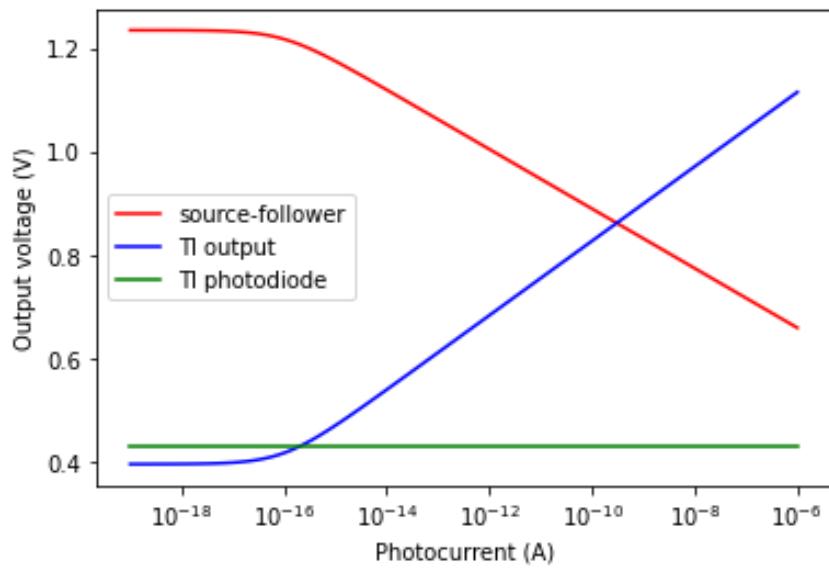


Figure 114: DC Photoreceptors responses

10.10.2 Large signal transient response

Next, we compute the large signal transient response. Similarly to the lab on linear systems, we need to compute the ordinary differential equation for both photo receptors.

Less photocurrent translates to more output voltage in the SF photoreceptor. In the time domain with a large input step, SF response is almost symmetrical to the current, whereas TI response behaves similar to a Follower Integrator.

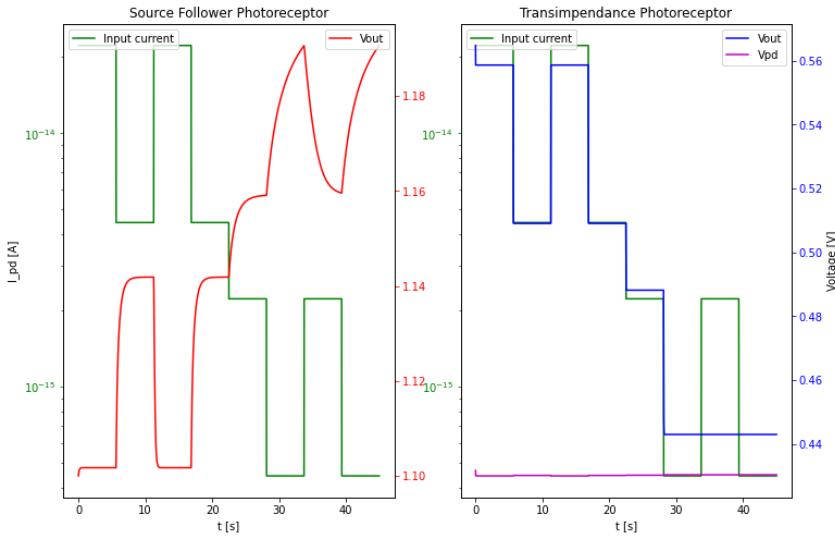


Figure 115: Large signal transient photoreceptors responses

10.10.3 Small signal transient response

Frequency response TI photoreceptor is quicker in response to changes in photocurrent than the SF photoreceptor (and also noisier).

In the simulation, we found a SF cutoff frequency of 289.31mHz and a TI cutoff frequency of 74.35Hz .

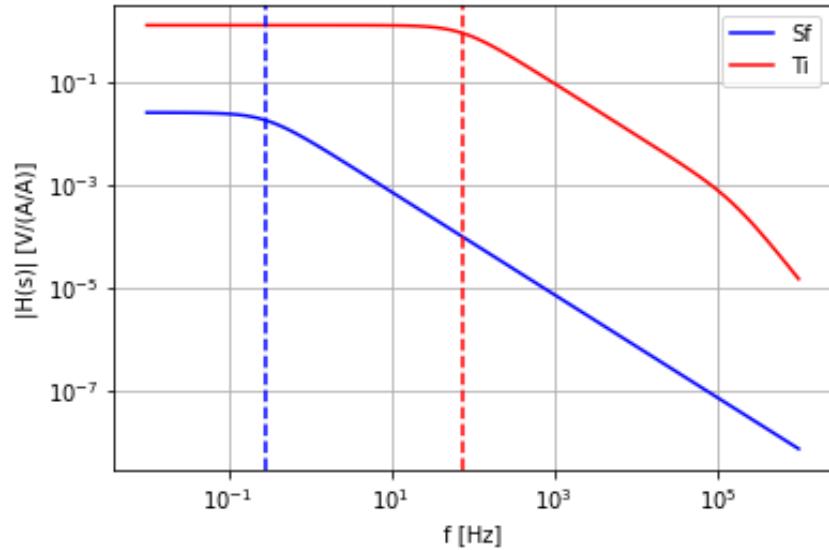


Figure 116: Bode plots of the photoreceptors

To explain this numbers, we referred to the feedback of the TI photoreceptor, which increases tau and decreases the frequency. Since there are no peaks in the frequency response of the bode plot, the TI circuit moves slowly toward equilibrium (overdamped) at this I_b and I_{pd} .

Root locus plot The pole of the SF photoreceptor is defined as the frequency for which the value of the denominator of its transfer function becomes zero (as I like to say the poles define the undefined).

The SF photoreceptor has a single pole and it moves farther away from the origin, as the photocurrent increases.

On the other hand, the TI photoreceptor has two poles, because of the quadratic demoninator $D(s)$ of $H(s)$. To study them and plot their locations on a complex plane, we need what are called root locus plots.

Root locus plots is a graphical analysis method used in control theory and stability theory. They examine how the roots of a system change with variation of a certain system parameter (in our case we will increase the amplifier bias current I_b).

The quality factor Q exists these poles.

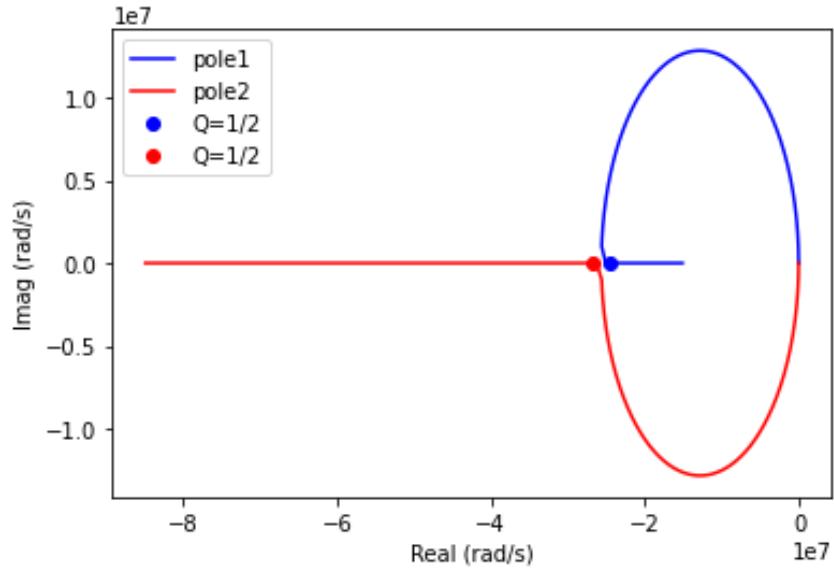


Figure 117: Poles of the transimpedance photoreceptor

In the exactly critically-damped condition, the quality factor Q equals 0.5 . This means that at around $Q = 0.5$, the voltage output of the TI photoreceptor returns as quickly as possible to its equilibrium position without oscillating back and forth.

Experiment From the conditions we defined above, we can experiment the small signal transient input photocurrent.

We define a small signal transient input photocurrent with the following step size.

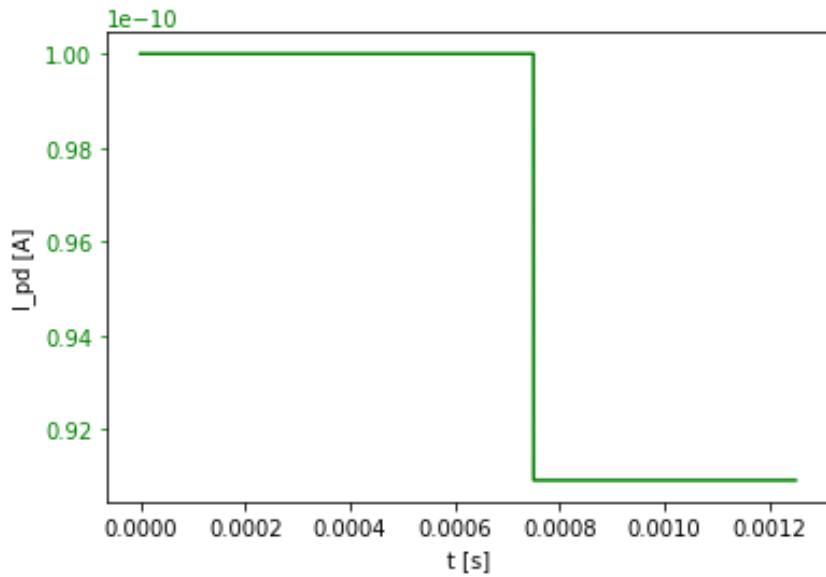


Figure 118: Small signal transient input photocurrent for TI photoreceptor

With a quality factor of about 1/2, the TI photoreceptor biais current smooth the small signal transient input photocurrent. Doubling Q , let the curve approach more closely the input.

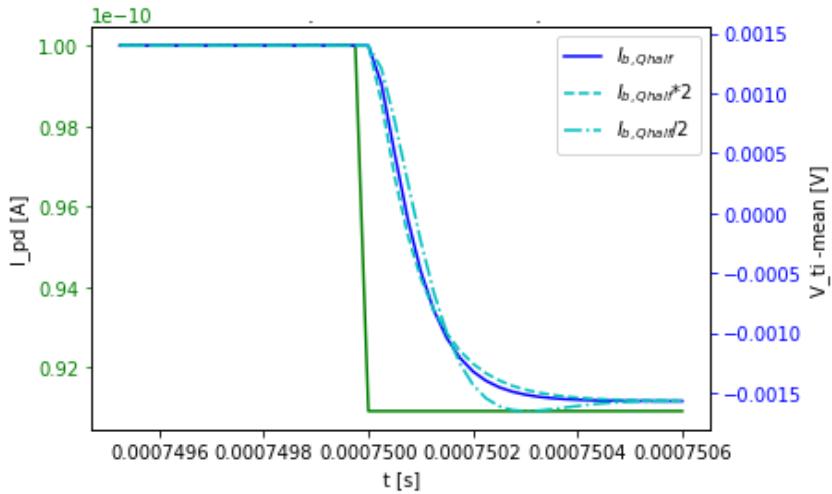


Figure 119: Small signal transient input photocurrent for TI photoreceptor, $Q = 0.5$

We found the theoretical maximum Q at 8.95, using a bias current of around $4e-10\text{A}$. However in such circumstances, the photoreceptor output has a ringing behavior. Apparently the poles of the transfer function are not purely real.

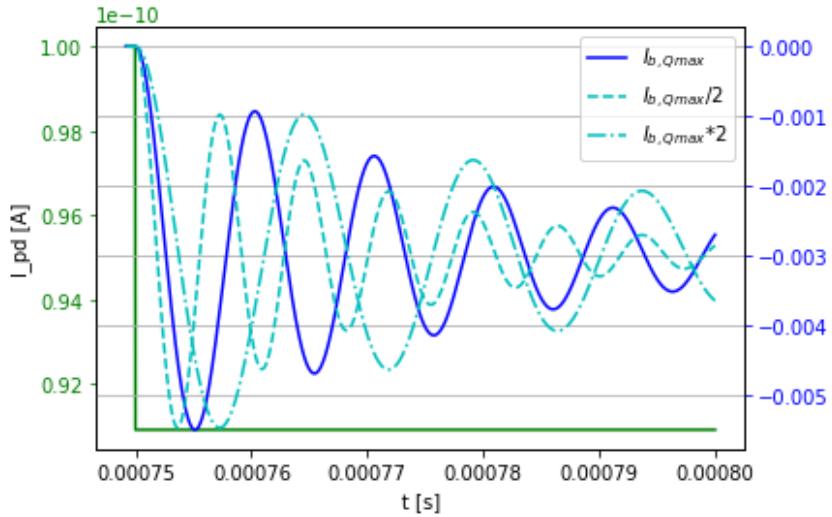


Figure 120: Small signal transient input photocurrent for TI photoreceptor, $Q = \text{max}$

10.11 Adaptive photoreceptor

Since biological photoreceptors are adaptive and amplify changes more than static inputs, a silicon photoreceptor that mimics this by providing a low gain for DC but a high gain for AC is the adaptive photoreceptor circuit given in 121.

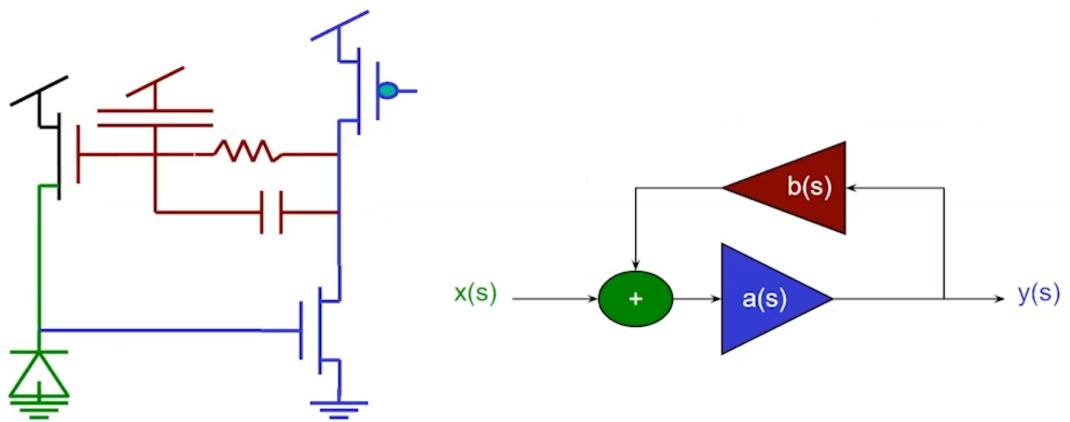


Figure 121: Adaptive photoreceptor, as shown in the lecture slides.

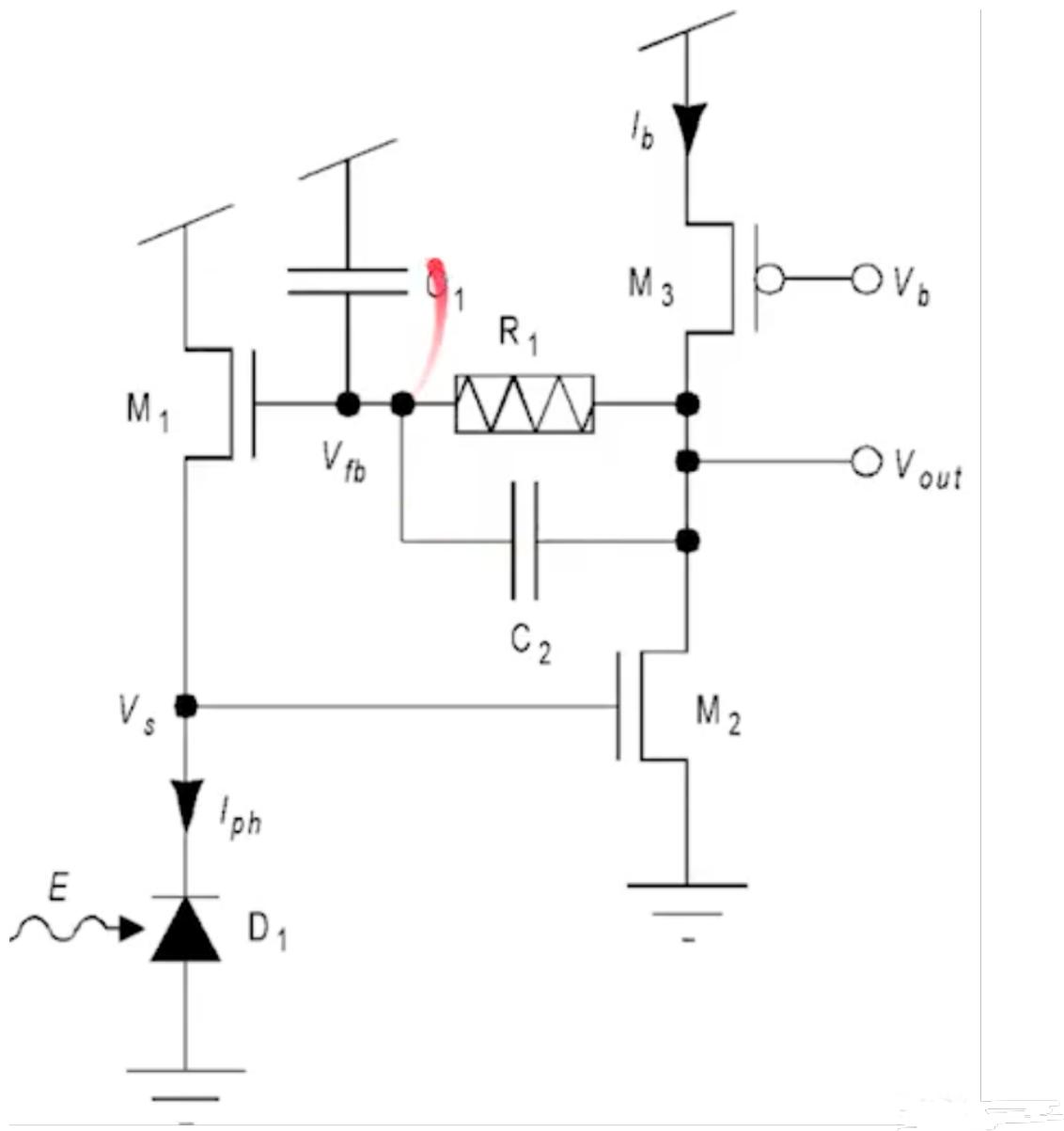


Figure 122: Adaptive photoreceptor in detail.

10.12 Cascode and Miller Effect

10.13 Test Yourself

This chapter being quite dense, you should expect general knowledge questions about the content.

Here is a list of things you should know about this chapter for the exam:

- What is the energy of a photon of a given wavelength?
- How does photo-transduction occur in silicon, i.e. how are electrons generated and how are they collected in a PN junction?
- What is Quantum Efficiency (QE)?
- How does the I-V curve of a diode change in the presence of light?
- How does absorption length of photons change with wavelength over the visible and near infrared range?

- How can you build a simple logarithmic photoreceptor?
- How you can use adaptation in a feedback loop to cancel out circuit mismatch.
- How you can build a fast logarithmic current-sense amplifier, by using feedback to make a virtual ground.
- How you can use a capacitive divider in the feedback loop of an amplifier to set a gain.
- How you can use a cascode configuration to increase effective drain resistance.
- What is the Miller effect?
- How can a cascode can be used to reduce it?

11 Complementary Notes and Topics

11.1 Revision Table

11.2 Know your constants - Tobi will ask you

Tobi will ask you to give constants. Yes I know, why would you bother remembering this in the age of internet? Well, general knowledge and showing off in dinner conversation is one reason. Truth is, it's important because it gives you an idea of the scale of things. Knowing the units is also very important, arguably more important than the numbers - helps you properly understand the purpose of the given constant.

Here are the constants you should know:

- *Kappa*
- Charge of an Electron: $e = -1.6 \cdot 10^{-19} \text{Coulomb}$
- Boltzmann Constant: $k = 1.38 \cdot 10^{-23} \text{J/K}$
- Thermal voltage: $U_T = \frac{kT}{q} = 25 \text{mV}$ at room temperature (T approx 300 Kelvin)
- Bandgap of Silicon: 1.12 eV
- Planck's Constant $h = 6.64 \cdot 10^{-34} \text{J.s}$
- Speed of light in vacuum: $c = 3 \cdot 10^8 \text{m.s}^{-1}$
- Highest photon wavelength that can generate an electron-hole pair in Silicon $\lambda_{max} = 1.1 \mu\text{m}$
- 1 Lux $\approx 10^4 \text{photons}/\mu\text{m}^2/\text{s}$
- Illumination at moonlight: 0.1 lux. Illumination with full sun: 100.000 lux

11.3 Exam Tips

11.3.1 General comments on the exam format and expected level of details

- We believe that we may be asked a series of related questions rather than unrelated questions of different chapters. For example, we could first be asked basic Saturation vs Ohmic transistor operation information, followed by two basic circuit such as the diff pair and the current mirror - which are the building blocks of the transconductance amplifier. We'd then of course be asked the transconductance amplifier. This logic can be applied to several different topics covered in this course.
- We think that general knowledge questions would also be asked. That is history of Neuromorphic Engineering, of transistors, Neuroscience topics, and physical constants or important quantities discussed in the course (U_T , κ , I_0 etc...)

11.3.2 Questions that they we think will be asked

- How do you measure κ if you don't have access to the physics of the devices?
You could use a NFET source follower, which transfer function includes Kappa, two voltages which one can control (from current source and the input transistor voltage) and one output voltage that you can measure.
- You could be given a basic circuit with example voltages and ask to derive the voltage/current of a given element, or simply say if everything makes sense or if we're violating some assumption like saturation.

11.3.3 How to practice your circuits

- Draw your circuits, and state all assumptions.
- Draw your circuits in opposite configuration (e.g. N and P type current mirror)
- Reason your way through circuits, with voltages/current starting point and everything that follows from this. This was done on the photodiode chapter. Everytime you can take assumptions such as κ_n and $\kappa_p = 1$, $I_{0n} = I_{0p}$, all transistors have the same W/L , neglect the Early Effect (except if important such as in small signal difference WTA for example).

12 Appendix

Neuromorphic Engineering I Study Guide

August 30, 2021

This is the collection of "What we expect you to know" sections from the individual labs. If you can tell someone else with confidence about these topics, you will do well at the exam. We recommend that you try to study with another student. Take turns explaining to each other about these topics, and drawing schematics and diagrams for each other.

1 Semiconductors

What is neuromorphic aVLSI all about? Who came up with it? Where is it being done?

Semiconductors, bands, band gap of silicon, the Fermi-Dirac distribution, doners and acceptors, PN junction, reverse and forward characteristics.

Using the Keithley voltage source (K230) and source-measure electrometer (K236) instruments. E.g. what is the K236 DC input impedance when measuring voltage with it? Basic knowledge of matlab. How to use the pot-boxes.

2 Subthreshold FET operation

What does it mean for a MOS transistor channel to be accumulated, flat-band, depleted, inverted? Knowledge of how subthreshold transistor operation is a diffusion process and why it depends exponentially on the terminal voltages. What is the meaning of "saturation"? What is the triode or linear operating range? I_{ds} vs V_{gs} on log scale. Differences between n- and p-fets. Typical values of I_0 , κ and subthreshold operating range. What are wells and how should the wells be biased relative to the substrate? What is the "back gate" or "body effect"? How is the back gate related to κ ? How to make a MOS capacitor and what is its C-V relationship. How a source follower works and how to compute the gain of a source follower.

3 Above-threshold FET operation

How transistors work above threshold. What is the linear or triode region and what is the saturation region? How do they depend on gate and threshold voltage. How the Early effect comes about. Typical values for Early voltage. How to sketch graphs of transistor current vs. gate voltage and drain-source voltage. How above-threshold transistors go into saturation and why the saturation voltage is equal to the gate overdrive. The above-threshold current equations. How above-threshold current depends on C_{ox} and mobility. How transconductance and drain resistance combine to generate voltage gain and what is the intrinsic voltage gain of a transistor. What effect does velocity saturation have on transistor operation? What is DIBL (drain induced barrier lowering) and II (impact ionization)? How transconductance and drain resistance combine to generate voltage gain.

4 Differential pairs, current correlator, bump circuits

Can you sketch a transamp, a wide range transamp, a current correlator, and a bump circuit in both n- and p-type varieties?

How does a differential pair work? How does the common-node voltage change with the input voltages? How can you compute the differential tail currents from the subthreshold equations, and how do you obtain the result in terms of the differential input voltage?

How does a current-correlator work? How does a bump circuit work?

The I-V characteristics of a transconductance amplifier below threshold. What's the functional difference between simple and wide-output-range transamp? The subthreshold transconductance g_m . The relation between gain A , transistor drain conductances g_d , and transconductances g_m .

Can you reason through all the node voltages in these circuits? I.e., if we draw the circuit and provide specific power supply and input voltages, can you reason to estimate all the other node voltages, at least to first order approximations, assuming $\kappa = 1$?

5 Transconductance amplifier

The I-V characteristics of a transconductance amplifier below threshold. How the open-circuit characteristics differ between simple and wide-output-range transamp. The subthreshold transconductance g_m . The relation between A , g_d , and g_m .

6 Followers, transimpedance amplifiers

How to use a non-linear circuit, such as a operational amplifier, to compute linear functions. What a unity-gain follower is used for. How to build a linear current-to-voltage converter (although we didn't build one in the lab).

7 Linear systems analysis, Time domain, follower integrator and differentiator, Hysteretic differentiator, Second-order systems

How to compute the time-constant of a low-pass filter and how to estimate it from the measurements. How to change the time-constant of a follower-integrator circuit.

The idea of using a lowpass filter (an integrator) to make a highpass filter (a differentiator). How the differentiator circuit is not a real differentiator but only an approximation over some frequencies defined by the time constant. How to sketch the transfer function of a differentiator circuit, showing the time constant on the sketch. How to implement a simple follower-differentiator.

How these circuits behave when driven with large signals.

8 Winner-take-all circuit (WTA)

How does the WTA circuit work? Can you reason through its behavior? How does the bias current affect its performance? How can you adjust the gain of the circuit by layout of the

transistors?

9 Photodiodes, photoreceptor

What is the energy of a photon of a given wavelength? How does phototransduction occur in silicon, i.e. how are electrons generated and how are they collected in a PN junction? What is QE? How does the I-V curve of a diode change in the presence of light? How does absorption length of photons change with wavelength over the visible and near infrared range? How can you build a simple logarithmic photoreceptor?

How you can use adaptation in a feedback loop to cancel out circuit mismatch. How you can build a fast logarithmic current-sense amplifier, by using feedback to make a virtual ground. How you can use a capacitive divider in the feedback loop of an amplifier to set a gain. How you can use a cascode configuration to increase effective drain resistance. What is the Miller effect? How can a cascode can be used to reduce it?

10 Neurons, axon-hillock circuit

What is a neuron and what are its components (synapse, soma, dendrite)? What types of models are used to simulate neurons? How does the spike-generating mechanism work? What is an FI curve? Can you draw the circuit schematic of the axon-hillock neuron?

11 Synapses, adaptive neuron circuit

The schematic for a synapse circuit. How the synaptic current changes as a function of the presynaptic frequency and the synaptic weight. How the firing frequency of an adaptive neuron changes as a function of the presynaptic frequency and the biases to the adapting synapse.

12 Learning in silicon: Tunneling, hot electron injection

How do tunneling and injection mechanisms work? What is the shape of the energy band diagram in the channel and oxide during tunneling and injection? How are the memory cell circuits used to control tunneling and injection?

13 Long range communication, Address-Event representation

What is AER? How does an AER system mimic biology's use of nerves? What constraints on firing rate are important for an AER system? What does it mean to be asynchronous? To be delay insensitive? What are the advantages and disadvantages of asynchronous designs? What is an isochronous fork assumption? What does the term bundled data mean? What is a dual rail encoding? What are the advantages and disadvantages of bundled data? How does a 4-phase handshake work? What is an arbiter? What is a C-element and can you draw a circuit for one? What is a staticizer and can you draw the circuit for one? Can you sketch the communication architecture and timing for a 2-d AER chip (like a retina)?

8 Linear Systems Theory

In this chapter we will review some properties of linear time-invariant systems. We consider their input/output relationship in the time domain, the impulse response, and the convolution theorem. We also review basic concepts of complex number theory; the Laplace transform, system's transfer function, and frequency domain analysis. More extended descriptions of this material can be found in many standard textbooks (Carlson, 1986; Pouliakis and Seely, 1994; Oppenheim et al., 1996).

8.1 Linear Shift-Invariant Systems

In linear systems theory, a system is treated as a *black box* that does not reveal its internal states, and is characterized only by the relationship between its input and output (see Fig. 8.1). If a system has no internal stored energy, then its output response $y(t)$ is forced entirely by the input $x(t)$:

$$y(t) = F[x(t)] \quad (8.1.1)$$

where $F[\cdot]$ is the transfer function.

Linearity A system is linear if it obeys the two fundamental principles: **Homogeneity**, and **additivity**.

The principle of homogeneity states that output scales linearly with the input:

$$F[\alpha x(t)] = \alpha F[x(t)]. \quad (8.1.2)$$

Usually, linear systems theory is applied to time-varying signals. However the same methods can be applied to input and output signals that are distributed

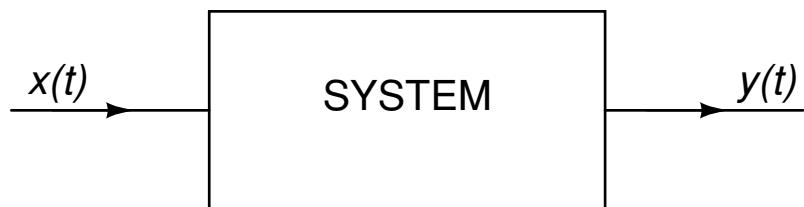
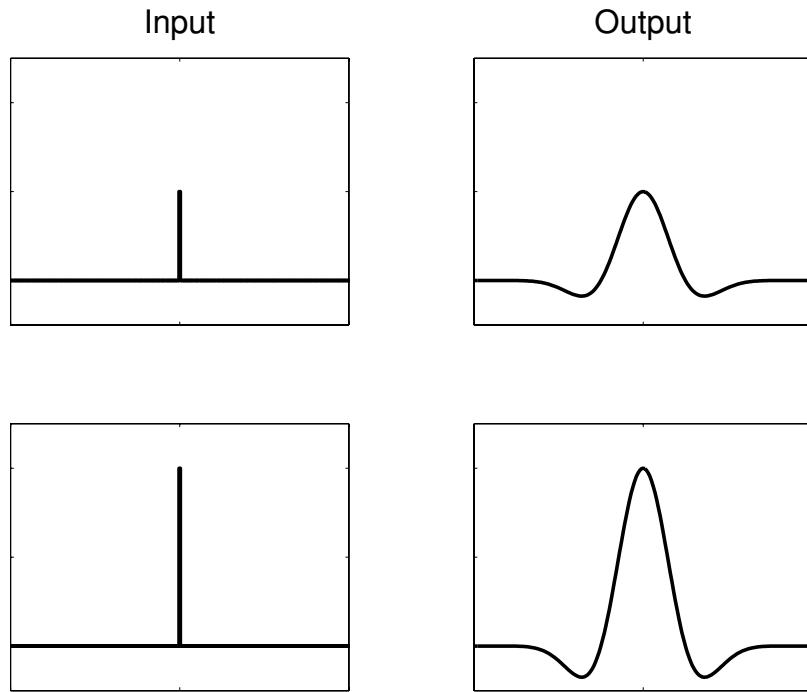


Figure 8.1

Typical black-box representation of a linear system. Its input is the signal $x(t)$ and its output is the signal $y(t)$.

**Figure 8.2**

Graphical example of the homogeneity principle of a linear system. The signals in the left quadrants represent the system's input, and the signals in the right ones represent its output. An increase in the input signal causes a proportional increase in the output signal.

over *space* rather than *time*. For example, in Fig. 8.2, the input signal is a spatial unit impulse and the output is a spatial *Gabor* function (a Gaussian modulated by a cosine function)¹

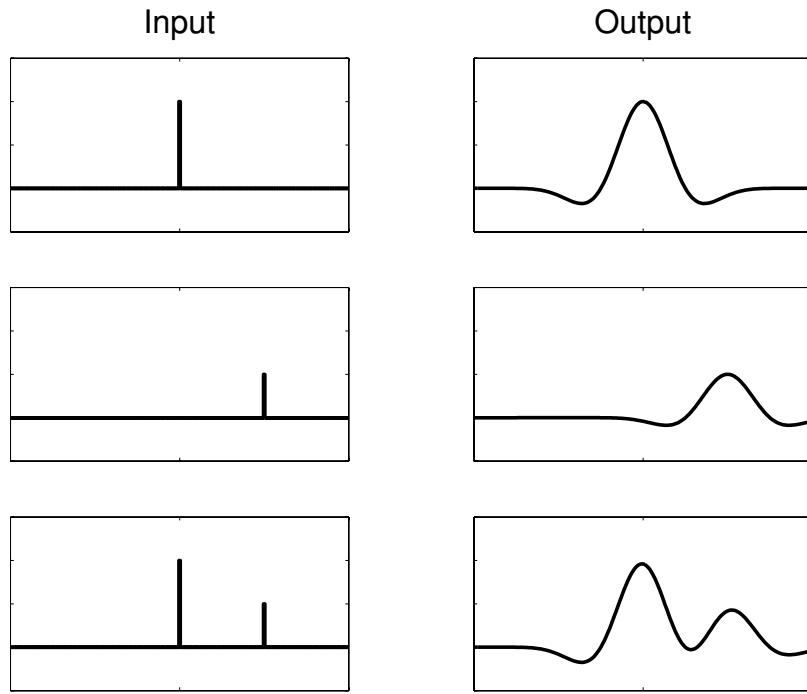
The principle of additivity states that if the input signal is composed of elementary signals, then the system's response is the composition of its responses to each of the elementary signals:

$$F[x_1(t) + x_2(t) + \dots + x_n(t)] = F[x_1(t)] + F[x_2(t)] + \dots + F[x_n(t)]. \quad (8.1.3)$$

Figure 8.3 shows a graphical example: If the response of the system to a spatial impulse is a Gabor function, and if the system's input signal is a linear combination of spatial unit impulses, then the system's response will be a linear combination of Gabor functions.

The principles of homogeneity and additivity taken together are commonly referred to as the *principle of superposition*, which state that a system is linear

¹ Gabor functions are commonly used to model the (linear) response properties of a particular class of neurons in the visual cortex.

**Figure 8.3**

Graphical example of the additivity principle of a linear system. The signals in the left quadrants represent the system's input, and the signals in the right ones represent its output.

if

$$y(t) = \sum_k a_k F[x_k(t)] \quad (8.1.4)$$

for input $x(t) = \sum_k a_k x_k(t)$, and a_k constant for all k .

In other words, a system is linear if its response function F is a *linear operator*:

$$F \left[\sum_k a_k x_k(t) \right] = a_k \sum_k F[x_k(t)]. \quad (8.1.5)$$

Shift Invariance A system is said to be shift-invariant if its responses to identical stimuli shifted in time are also identical, except for the corresponding time shift (Fig. 8.4).

If a system is shift-invariant, then its response function is also shift-invariant: Given input signal $x(t)$, its time-shifted variant $x(t - \tau)$ will produce

$$F[x(t - \tau)] = y(t - \tau). \quad (8.1.6)$$

The system's output signal is unchanged, except for a time shift.

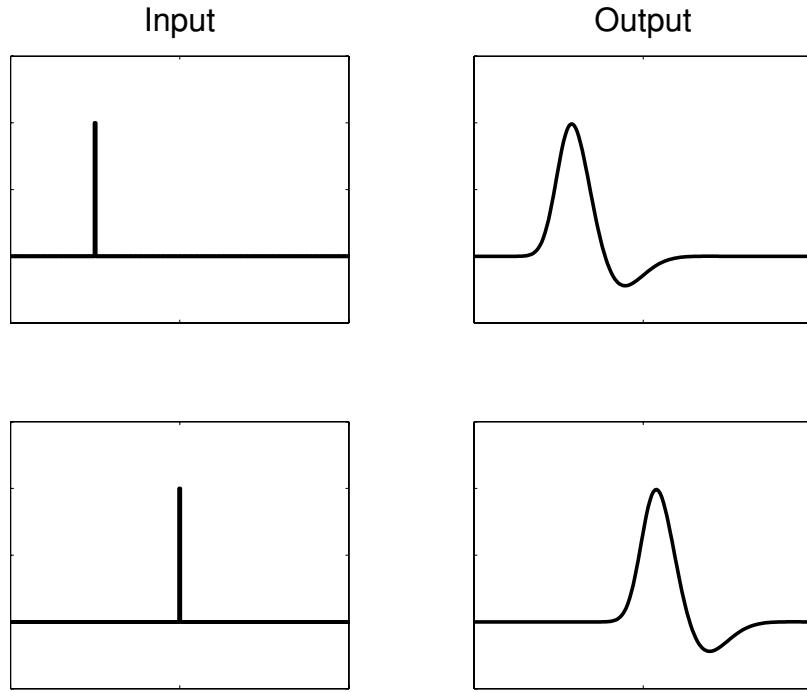


Figure 8.4
Graphical example of a time-invariant system's response.

Time invariance and linearity are two independent characteristics. *Not all linear systems are time-invariant and, similarly, not all time-invariant systems are linear.*

8.2 Convolution

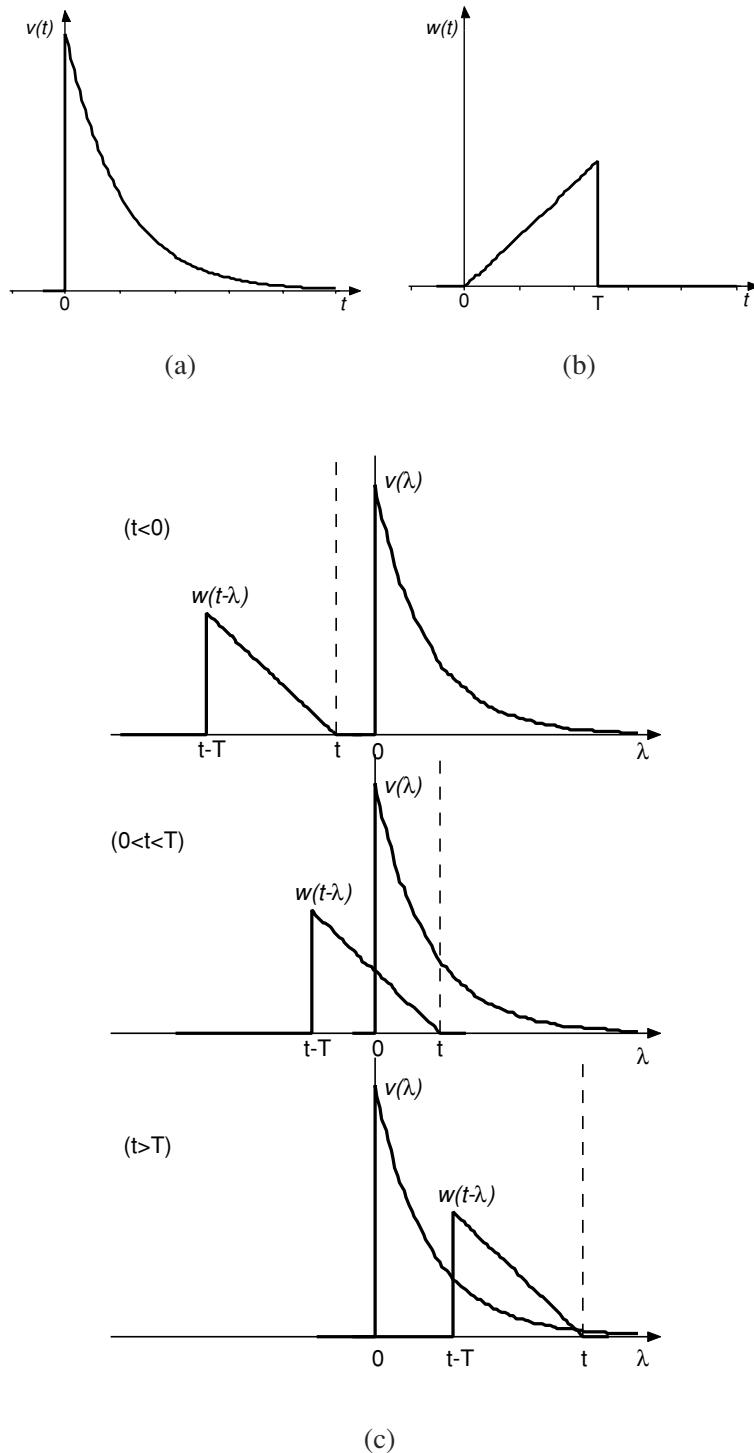
Convolution is an important mathematical operator used in linear systems analysis. The convolution of two time-varying signals, $v(t)$ and $w(t)$, is

$$v(t) * w(t) \equiv \int_{-\infty}^{+\infty} v(\lambda)w(t - \lambda)d\lambda \quad (8.2.1)$$

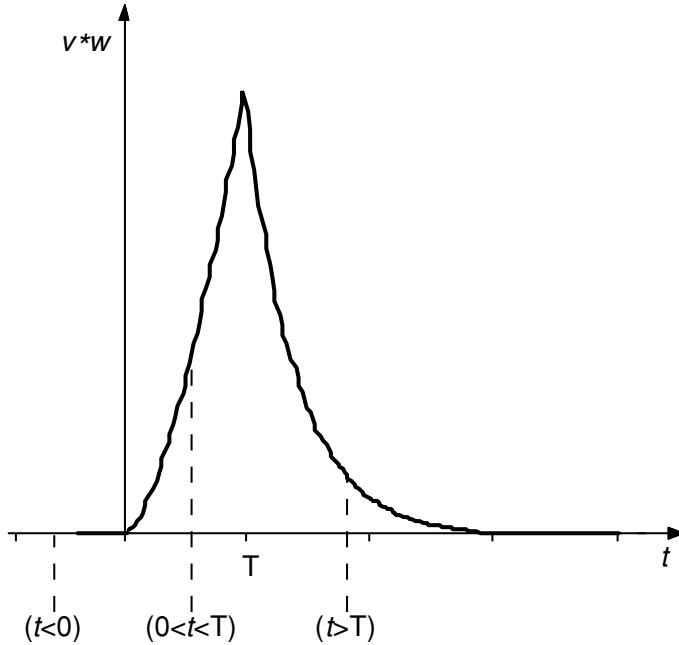
where λ is the integration variable, and t is the independent variable.

Figure 8.5 shows a graphical representation of the convolution process between signals $v(t)$ (Fig. 8.5(a)) and $w(t)$ (Fig. 8.5(b)) at three different time steps. The result of the convolution is shown in Fig. 8.6. Note how the overlap between the two curves is null for $t < 0$, increases for $0 < t < T$, peaks at $t = T$ and decreases again for $t > T$.

If the independent variable for both input signals is the same, then it can be omitted and we express the convolution between the two signals $v(t)$ and

**Figure 8.5**

Graphical representation of the convolution between $v(t)$ and $w(t)$ for three different values of t . Note that the integration variable λ in (c) is on the abscissae of the plots. Modified from Carlson, A. B. (1986).

**Figure 8.6**

Result of the convolution between the two signals $v(t)$ and $w(t)$ of Fig. 8.5. The three dashed lines are at the three values of t used in Fig. 8.5.

$w(t)$ simply as $v * w$. The convolution operator is linear and has the following properties:

commutative:	$v * w$	$=$	$w * v$
associative:	$v * (w * z)$	$=$	$(v * w) * z$
distributive:	$v * (w + z)$	$=$	$(v * w) + (v * z)$

8.3 Impulses

The *unit impulse* or *Dirac delta function* $\delta(t)$ is not a function in the strict mathematical sense. It is defined by a set of assignment rules.

- If $v(t)$ is a continuous function at $t = 0$ then

$$\int_{t_1}^{t_2} v(t)\delta(t)dt = \begin{cases} v(0) & t_1 < 0 < t_2 \\ 0 & \text{otherwise.} \end{cases} \quad (8.3.1)$$

- If ϵ is an arbitrary small number

$$\int_{-\infty}^{+\infty} \delta(t)dt = \int_{-\epsilon}^{+\epsilon} \delta(t)dt = 1. \quad (8.3.2)$$

From these rules we can infer that $\delta(t)$ has unit area at $t = 0$ and that $\delta(t) = 0$, for all $t \neq 0$. We can also note that the Dirac delta function has no mathematical or physical meaning, unless it appears under the integral operator.

Impulse Integration Properties

When used in conjunction with the integral operator, the Dirac delta function has the following properties:

- Replication:

$$v(t) * \delta(t - \tau) = v(t - \tau) \quad (8.3.3)$$

- Sampling:

$$\int_{-\infty}^{+\infty} v(t)\delta(t - \tau)dt = v(\tau) \quad (8.3.4)$$

where $v(t)$ is a continuous time-varying signal.

Impulses in the Limit

There are many (proper mathematical) functions $\delta_\epsilon(t)$ that approach the Dirac delta function $\delta(t)$, in the limit:

$$\lim_{\epsilon \rightarrow 0} \delta_\epsilon(t) = \delta(t) \quad (8.3.5)$$

An example of such a function that is commonly used is

$$\delta_\epsilon = \frac{\sin(\frac{t}{\epsilon})}{t}. \quad (8.3.6)$$

Figure 8.7 shows how δ_ϵ of Eq. 8.3.6 approaches the Dirac delta function as ϵ decreases.

8.4 Impulse Response of a System

We can now use the notions of convolution and unit impulse to define the *impulse response* of a linear time-invariant system. If $y(t)$ is the system's response to its input $x(t)$ we can write

$$y(t) = F[x(t)]. \quad (8.4.1)$$

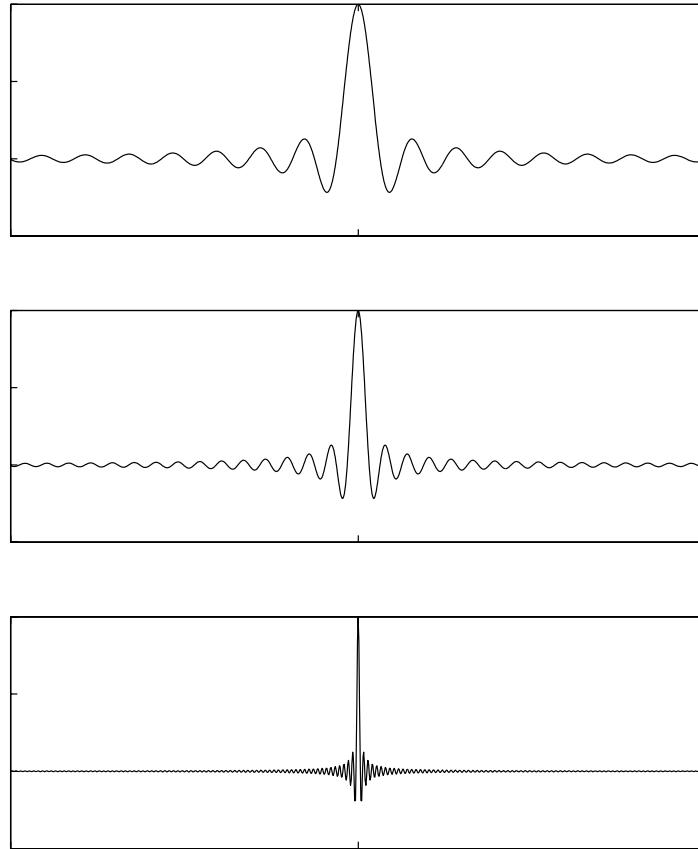


Figure 8.7
Plot of the function $\sin(t/\epsilon)/t$ for three decreasing values of ϵ .

If the input signal is the Dirac delta function ($x(t) = \delta(t)$), then the system's response to the unit impulse is defined as

$$h(t) \equiv F[\delta(t)]. \quad (8.4.2)$$

If $x(t)$ is continuous in time, the replication property of the unit impulse allows us to rewrite $x(t)$ as $x(t) * \delta(t)$. With this reformulation of the system's input signal, Eq. 8.4.1 becomes

$$y(t) = F \left[\int_{-\infty}^{+\infty} x(\lambda) \delta(t - \lambda) d\lambda \right]. \quad (8.4.3)$$

If the system is linear, Eq. 8.4.3 is equivalent to:

$$y(t) = \int_{-\infty}^{+\infty} x(\lambda) F[\delta(t - \lambda)] d\lambda \quad (8.4.4)$$

If we substitute Eq. 8.4.2 into Eq. 8.4.4, and if the system is time-invariant,

then

$$y(t) = \int_{-\infty}^{+\infty} x(\lambda)h(t - \lambda)d\lambda = \int_{-\infty}^{+\infty} h(\lambda)x(t - \lambda)d\lambda. \quad (8.4.5)$$

This property of linear time-invariant systems is extremely powerful. It states that if a system's impulse response $h(t)$ is known, the response of the system to any arbitrary signal $x(t)$ can be computed simply by performing the convolution of its impulse response with the signal itself:

$$\boxed{y(t) = h(t) * x(t)}. \quad (8.4.6)$$

Step Response

We can define a system's *step response* in the same way we defined its impulse response. If the input signal $x(t)$ is the step function

$$u(t) = \begin{cases} 1 & \text{if } t \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (8.4.7)$$

then we define its step response to be

$$g(t) \equiv F[u(t)]. \quad (8.4.8)$$

A first interesting property can be obtained by exploiting the system's impulse response (see Eq. 8.4.6):

$$g(t) = h(t) * u(t). \quad (8.4.9)$$

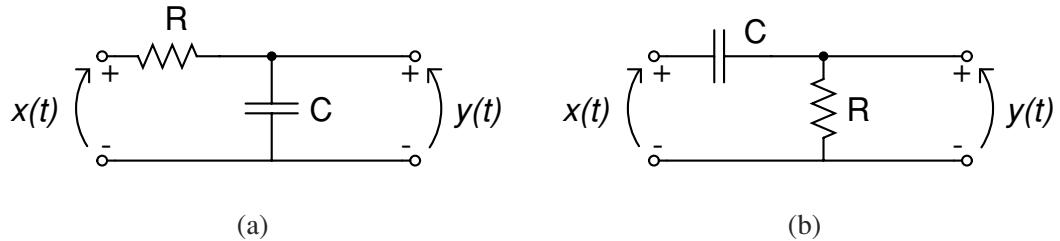
By applying the derivative operator to this equation, and noting that the derivative of the step function is the unit impulse, we obtain

$$\frac{d}{dt}g(t) = h(t) * \frac{d}{dt}u(t) = h(t) * \delta(t). \quad (8.4.10)$$

If we apply the unit impulse replication property (Eq. 8.3.3), then we obtain

$$h(t) = \frac{d}{dt}g(t). \quad (8.4.11)$$

Thus, a system's impulse response can be obtained by computing the derivative of its step response. This property is extremely useful in practical situations because unit impulses are impossible to generate with physical instruments but it is easy to generate waveforms that approximate ideal step functions. Consequently, a physical linear time-invariant system is characterized experimentally

**Figure 8.8**

Resistor capacitor (RC) circuits. The signals $x(t)$ represent input voltages, and the signals $y(t)$ represent output voltages. (a) Integrator circuit; (b) Differentiator circuit.

by measuring its step response and then deriving its impulse response from Eq. 8.4.11.

8.5 Resistor-Capacitor Circuits

The resistor-capacitor (RC) circuits of Fig. 8.8 represent first order, linear, time-invariant systems. In both circuits, the input signal is $x(t)$ and the output signal is $y(t)$. The circuits of Fig. 8.8(a) and (b) are referred to as *RC integrator* and *RC differentiator* respectively. In this section we focus only on the properties of the RC integrator. The properties of the RC differentiator will be described in Chapter 9.

The integrator circuit of Fig. 8.8(a) is governed by the differential equation:

$$RC \frac{d}{dt}y(t) + y(t) = x(t). \quad (8.5.1)$$

By solving Eq. 8.5.1 for a unit impulse input signal ($x(t) = \delta(t)$), we obtain the circuit's *impulse response*:

$$h(t) = \frac{1}{RC} e^{-t/RC} \cdot u(t) \quad (8.5.2)$$

where $u(t)$ is the step function. Similarly, solving Eq. 8.5.1 for a step input signal ($x(t) = u(t)$), we obtain the circuit's *step response*

$$g(t) = (1 - e^{-t/RC}) \cdot u(t). \quad (8.5.3)$$

Figure 8.9 shows the impulse response and the step response. The value RC is defined as the system's *time-constant* and is often labeled τ . As pointed out in Section 8.4, the response of the circuit to an arbitrary input signal can be

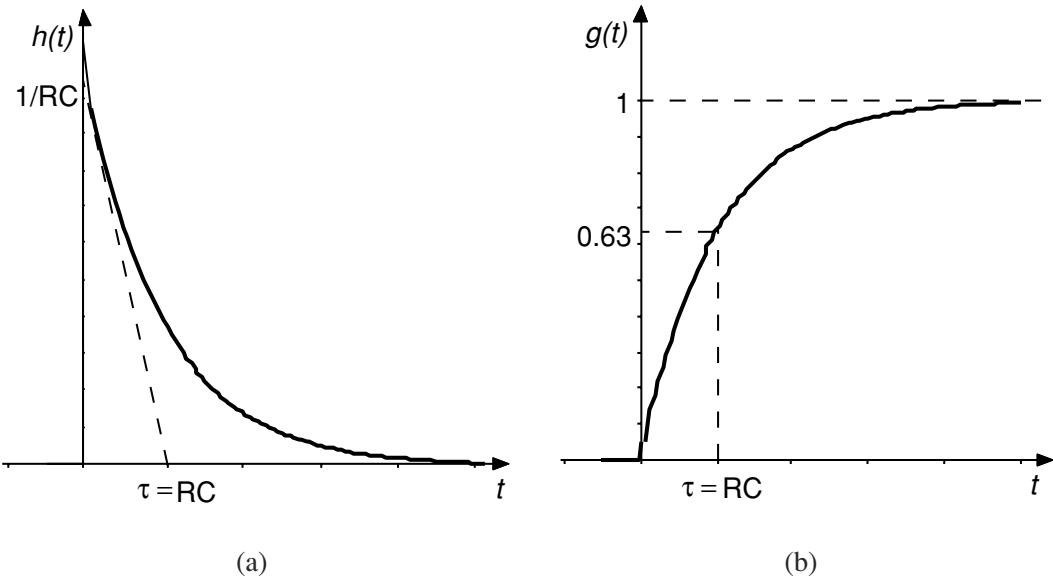


Figure 8.9
Impulse response (a) and step response (b) of an RC circuit.

obtained by the convolution between the input signal and the circuit's impulse response:

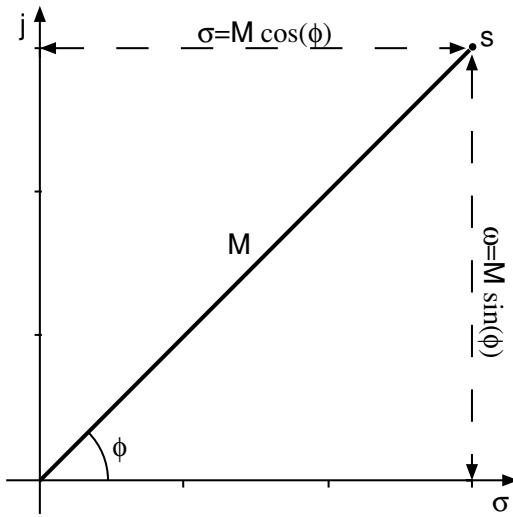
$$y(t) = x(t) * h(t) = \int_0^{\infty} \frac{1}{RC} e^{-\lambda/RC} \cdot x(t - \lambda) d\lambda. \quad (8.5.4)$$

8.6 Higher Order Equations

Time-domain analysis becomes increasingly difficult for higher order systems. Fortunately there is a *unified representation* in which any solution to a linear system can be expressed: **Exponentials with complex arguments**. All solutions to linear homogeneous (undriven) equations are of the form e^{st} where s is a *complex number* (see Fig. 8.10):

$$s = \sigma + j\omega = M \cos(\phi) + jM \sin(\phi) \quad (8.6.1)$$

where $j = \sqrt{-1}$, σ is the real part of the complex number, ω is the imaginary part, M represents its *magnitude*, and ϕ its *phase*. Magnitude and phase of a

**Figure 8.10**

Complex number representation. The complex number s has magnitude M and phase ϕ . Its real part is σ and imaginary part is ω .

complex number obey the following relationships:

$$M = \sqrt{\sigma^2 + \omega^2} \quad (8.6.2)$$

$$\phi = \arctan\left(\frac{\omega}{\sigma}\right). \quad (8.6.3)$$

The magnitude of a complex number s is often denoted as $|s|$. Furthermore, applying the properties of complex exponentials, one can observe that

$$e^{j\phi} = \cos(\phi) + j \sin(\phi) \quad (8.6.4)$$

$$e^{-j\phi} = \cos(\phi) - j \sin(\phi). \quad (8.6.5)$$

It follows that s can be also written as

$$s = M e^{j\phi}. \quad (8.6.6)$$

These notations can be used to solve higher order differential equations. As an example, we consider the second order linear homogeneous equation

$$\frac{d^2}{dt^2} V + \alpha \frac{d}{dt} V + \beta V = 0. \quad (8.6.7)$$

Assume that e^{st} is an *eigenfunction*² and substitute for V :

$$s^2 e^{st} + \alpha s e^{st} + \beta e^{st} = 0. \quad (8.6.8)$$

Solving for s we obtain

$$s = \frac{-\alpha \pm \sqrt{\alpha^2 - 4\beta}}{2}. \quad (8.6.9)$$

Consequently, if $\alpha^2 - 4\beta \geq 0$, s is real, otherwise s is a complex number. In practice, if Eq. 8.6.7 is a linear system, we could measure its response $V = e^{st}$ with a *real* instrument: but if e^{st} was a complex exponential, we would measure only its real component:

$$\operatorname{Re}\{e^{st}\} = \operatorname{Re}\left\{e^{(\sigma+j\omega)t}\right\} = e^{\sigma t} \operatorname{Re}\{e^{j\omega t}\}. \quad (8.6.10)$$

So the measured response of the system would be

$$V_{meas} = e^{\sigma t} \cos \omega t. \quad (8.6.11)$$

Figure 8.11 shows the possible kinds of response of V_{meas} for different values of ω and σ . If $\sigma < 0$ all the solutions are stable and decay to zero with time. If $\sigma > 0$ all solutions are unstable and diverge with time. If $\sigma = 0$ the solutions are naturally stable (they neither decay, nor diverge). All physical *passive* linear systems will have stable solutions ($\sigma < 0$). The ω axis scales the oscillation frequency f of a solution ($\omega = 2\pi f$).

8.7 The Heaviside-Laplace Transform

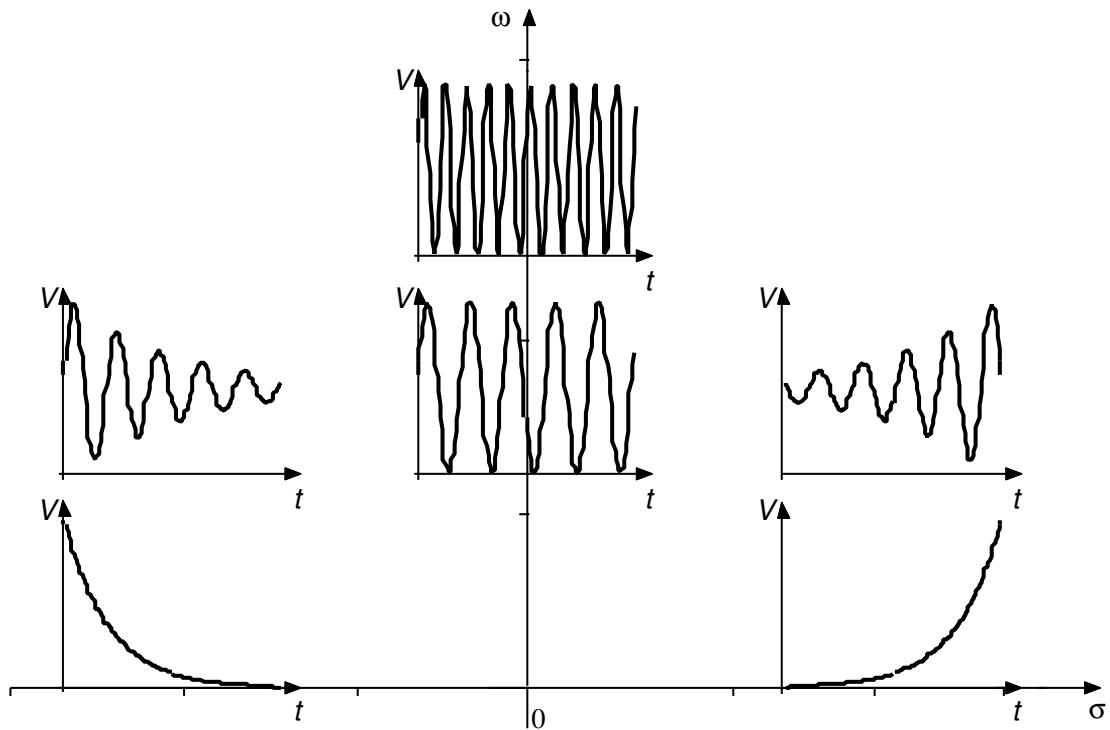
By analyzing the example of the previous section (see Eq. 8.6.7) we can make the following observation: Any time we substitute the eigenfunction e^{st} into a linear differential equation of order n , the following property obtains:

$$\frac{d^n}{dt^n} e^{st} = s^n e^{st}. \quad (8.7.1)$$

In other words:

We can consider s as an operator meaning *derivative* with respect to time. Similarly, we can view $\frac{1}{s}$ as the operator for *integration* with respect to time (Heaviside).

² An eigenfunction is a nonzero solution of a second order linear homogenous differential equation

**Figure 8.11**

The possible kinds of *measured* responses for a first order linear system.

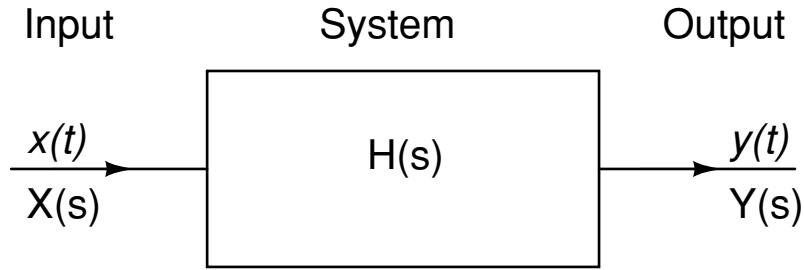
This observation was made by Heaviside, when trying to analyze analog circuits: but was also formalized by Laplace when he introduced the *Laplace Transform*. The Laplace transform is a useful operator that links functions that operate in the time domain with functions of complex variables:

$$\mathcal{L}[y(t)] = Y(s) \equiv \int_{-\infty}^{\infty} y(t)e^{-st}dt. \quad (8.7.2)$$

8.8 Linear System's Transfer Function

Now that we have introduced the concepts of convolution (Section 8.2), impulse response (Section 8.4), and the Laplace transform (Section 8.7), we can define a linear system's *transfer function*. It is a function defined in the complex domain:

$$H(s) \equiv \frac{Y(s)}{X(s)} \quad (8.8.1)$$

**Figure 8.12**

Typical representation of a linear system with input and output signals both in the time domain ($x(t), y(t)$) and in the Laplace domain ($X(s), Y(s)$).

where $Y(s)$ is the Laplace transform of the system's output $y(t)$ and $X(s)$ is the Laplace transform of the system's input $x(t)$ (see Fig. 8.12). Conversely, we can say that the output of any linear time-invariant system is determined by *multiplying* the system's transfer function with its input:

$$\boxed{Y(s) = H(s)X(s)}. \quad (8.8.2)$$

Transfer Function and Impulse Response

Consider the special case in which the system's input signal $x(t)$ is the unit impulse $x(t) = \delta(t)$. Its Laplace transform $X(s)$ is

$$X(s) = \int_{-\infty}^{\infty} x(t)e^{-st}dt = \int_{-\infty}^{\infty} \delta(t)e^{-st}dt = 1. \quad (8.8.3)$$

In this case, following the definition of Eq. 8.8.1, the system's response in the complex plane is

$$Y(s) = H(s). \quad (8.8.4)$$

On the other hand, the system's response in the time domain is (by definition) its impulse response:

$$y(t) = h(t). \quad (8.8.5)$$

Because $Y(s)$ is the Laplace transform of $y(t)$, we can substitute Eq. 8.8.5 into Eq. 8.7.2:

$$Y(s) = \int_{-\infty}^{\infty} y(t)e^{-st}dt = \int_{-\infty}^{\infty} h(t)e^{-st}dt \quad (8.8.6)$$

and so

$$H(s) = \int_{-\infty}^{\infty} h(\lambda)e^{-\lambda s} d\lambda = \mathcal{L}[h(t)]. \quad (8.8.7)$$

The transfer function $H(s)$ is the Laplace transform of the impulse response $h(t)$.

Summary Given a linear time-invariant system with input $x(t)$, output $y(t)$, and impulse response $h(t)$:

$$\begin{aligned} y(t) &= x(t) * h(t) \\ Y(s) &= X(s)H(s) \end{aligned}$$

where

$$\begin{aligned} X(s) &= \mathcal{L}[x(t)] \\ Y(s) &= \mathcal{L}[y(t)] \\ H(s) &= \mathcal{L}[h(t)]. \end{aligned}$$

8.9 The Resistor-Capacitor Circuit (A Second Look)

Consider again the RC circuit of Fig. 8.8. As mentioned in Section 8.5, this circuit is governed by

$$\tau \frac{d}{dt}y(t) + y(t) = x(t) \quad (8.9.1)$$

where $\tau = RC$.

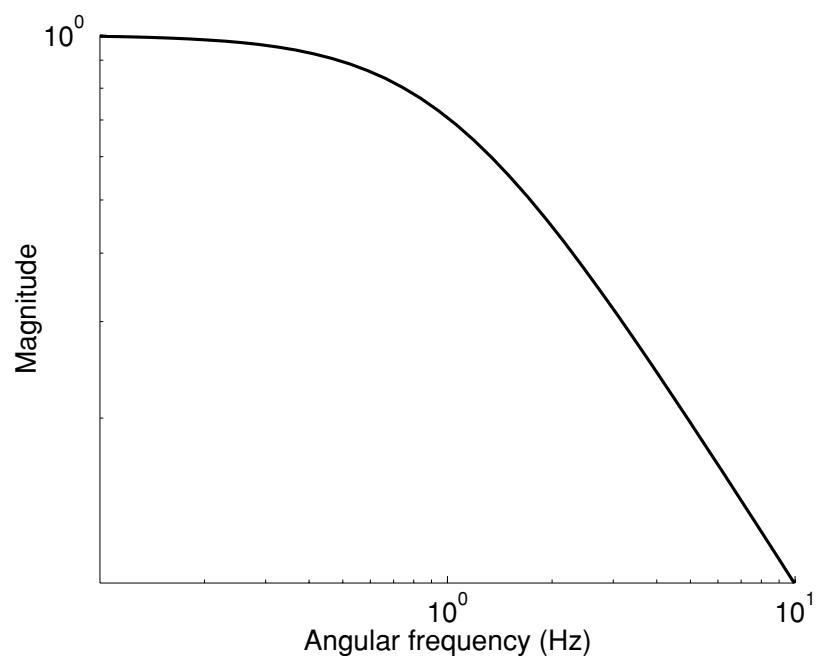
In the complex domain, we have

$$Y(s)(\tau s + 1) = X(s). \quad (8.9.2)$$

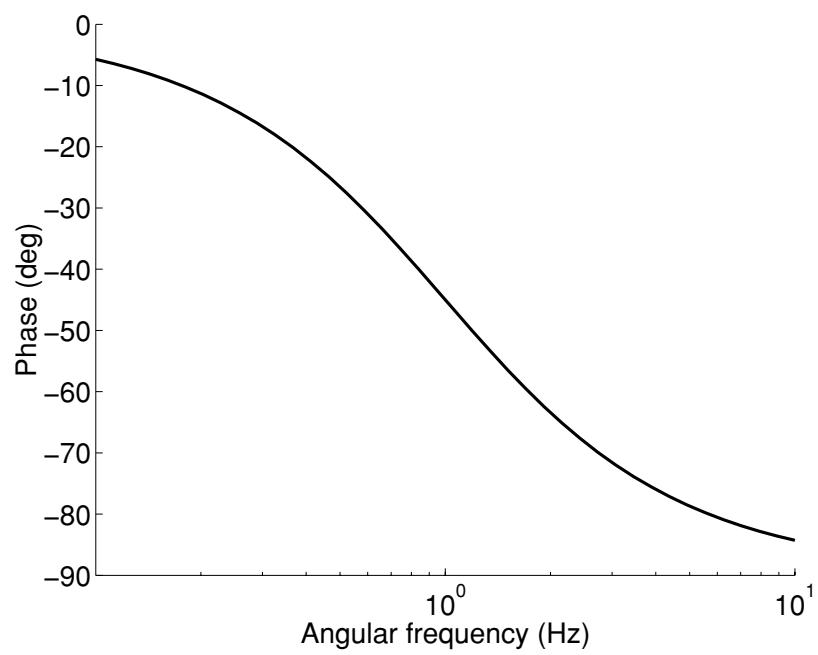
Therefore the circuit's transfer function is

$$H(s) = \frac{1}{1 + \tau s}. \quad (8.9.3)$$

Consider now how this circuit responds to sinusoidal signals of different *frequencies*. Sinusoids have a very special relationship to shift-invariant linear systems, such as the one we are analyzing. When a sinusoidal signal is applied



(a)



(b)

Figure 8.13

Bode plot of a first order linear system, such as the RC circuit of Fig. 8.8. (a) Magnitude, (b) Phase.

as input to a shift-invariant linear system, then its response will be another sinusoidal signal, with possibly a different amplitude and a different phase, but certainly with exactly the same frequency! That is, if the input is $x(t) = \sin(\omega t)$, the output will be $y(t) = A \sin(\omega t + \phi)$, where A and ϕ determine the scaling and shift.

When we analyze a system using sinusoidal signals of different frequencies, we are working in the frequency domain. In this domain $s = j\omega$ and the circuit's transfer function is

$$H(j\omega) = \frac{1}{1 + j\omega\tau}. \quad (8.9.4)$$

From this transfer function, we make two useful observations:

1. If the frequencies of the sinusoidal signals are small with respect to the circuit's time-constant ($\omega\tau \ll 1$), then the circuit's output will resemble its input ($Y(j\omega) \approx X(j\omega)$).
2. On the other hand, if the frequencies are large with respect to the circuit's time-constant ($\omega\tau \gg 1$), then

$$\frac{Y(j\omega)}{X(j\omega)} \approx \frac{1}{j\omega\tau}. \quad (8.9.5)$$

These observations are also reflected in the plots of the transfer function's magnitude and phase (Fig. 8.13). These plots are referred to as *Bode* plots and they are used to analyze the response of a dynamic system in terms of its transfer function. The magnitude of the transfer function is

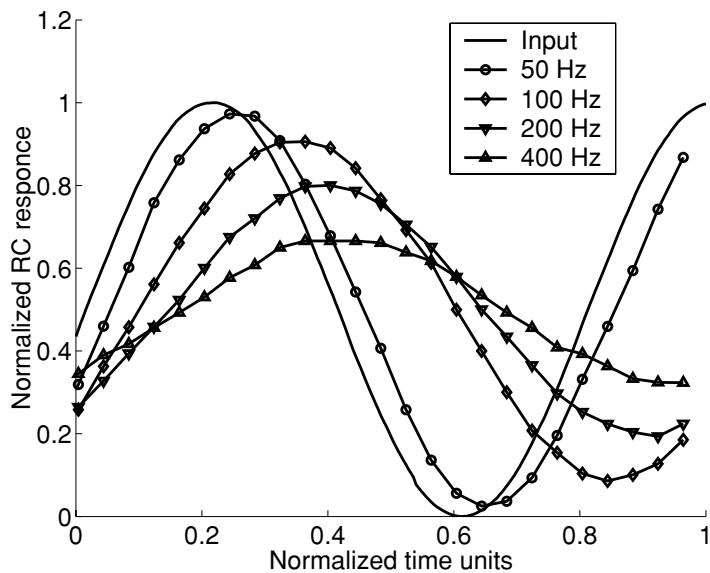
$$|H(j\omega)| = \frac{1}{\sqrt{1 + (\omega\tau)^2}} \quad (8.9.6)$$

and its phase is

$$\phi = \arctan(-\omega\tau). \quad (8.9.7)$$

The frequency $\omega = \frac{1}{\tau}$ is defined as the *cutoff frequency*. In Fig. 8.13 it is set to one.

The RC circuit of Fig. 8.8 is a *low-pass filter*, because it allows sinusoidal signals with frequencies lower than the cutoff frequency to pass virtually unchanged. On the other hand, the frequency components of the input signals that are above the cutoff frequency are attenuated. The phase lag between the input and the output of the system increases with ω (see Fig. 8.13(b)) and saturates at -90° . Figure 8.14 shows experimental data measured from an RC

**Figure 8.14**

Response of an RC low-pass filter ($R = 10M\Omega$, $C = 1nF$) to input sinusoids of different frequencies. The input signals have been normalized to unity, and the outputs have been normalized with respect to the input. The time axis has also been normalized so that the responses to all the frequencies could be presented on the same graph.

lowpass filter with $R = 10M\Omega$ and $C = 1nF$. Sinusoids of increasing frequency were applied to the circuit and the corresponding responses were measured. To show the effect of a range of input frequencies on the circuit's response, all the data are plotted on a normalized scale. The responses have been normalized with respect to the input and time has been normalized to unity. As expected, the output signal is attenuated as the input frequency increases; and the phase lag between the input and output signals increases with increasing frequency.

8.10 Low-Pass, High-Pass, and Band-Pass Filters

The RC circuit analyzed in the previous sections is the simplest example of a passive *filter*. Filters are typically used to alter the frequency spectrum of their input signals. Specifically, filters allow one or more frequency *bands* to pass unchanged (except for a multiplicative gain factor), whereas others are attenuated. Passive filters do not amplify the input signal, whereas active filters can also amplify the frequency components of the input signal. If a filter transmits low frequency components (from DC to a lower cutoff value ω_L), it is said to be a *low-pass* filter. If it transmits high frequency components (higher

than a cutoff value ω_u), it is said to be a *high-pass* filter. Filters that transmit only frequency components between a lower cutoff and an upper cutoff are said to be *band-pass* filters.

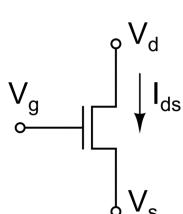
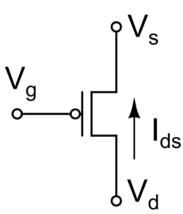
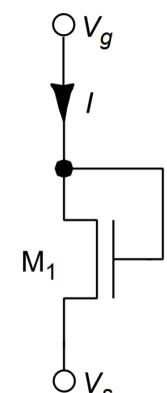
The analysis of linear systems often assumes *ideal filters*. These filters have distortionless signal transmission over one or more frequency bands, and have zeros responses at all other frequencies. For example, the transfer function of an ideal *bandpass* filter is:

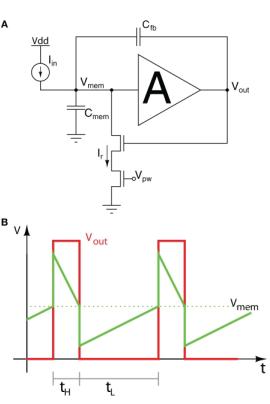
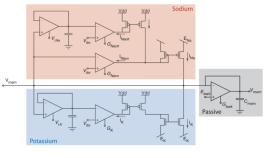
$$H(\omega) = \begin{cases} Ke^{-j\omega t_d} & \omega_l \leq |\omega| \leq \omega_u \\ 0 & \text{otherwise.} \end{cases} \quad (8.10.1)$$

Although ideal filters cannot be implemented in practice, their use in theoretical analysis simplifies the study of linear systems.

References

- [Com+20] Iulia M. Comsa **and others**. “Temporal Coding in Spiking Neural Networks with Alpha Synaptic Function: Learning with Backpropagation”. in *arXiv:1907.13223 [cs, q-bio]*: (november 2020). arXiv: 1907.13223. URL: <http://arxiv.org/abs/1907.13223> (urlseen 17/01/2022).

Circuit Diagram	Name, Objectives and Function	Key Equations	Assumptions
	NFET Current Source: Control current in a device through applied voltage.	In Saturation: $I_{ds} = I_{n0} e^{\frac{\kappa_n V_g - V_s}{U_T}}$ Not in Saturation: $I_{ds} = I_{n0} e^{\frac{\kappa_n V_g}{U_T}} \left(e^{\frac{-V_s}{U_T}} - e^{\frac{-V_d}{U_T}} \right)$	<ul style="list-style-type: none"> • Neglecting Early Effect (in Saturation) • Neglecting Back Gate Effect • Assuming $W/L = 1$ • Operating in Subthreshold • Probably a lot more assumptions • We'll name these assumptions <i>Standard Subthreshold Transistor Assumption</i>.
	PFET Current Source: Control current in a device through applied voltage.	In Saturation: $I_{sd} = I_{p0} e^{\frac{-\kappa_p V_g + V_s}{U_T}}$ Not in Saturation: $I_{sd} = I_{p0} e^{\frac{-\kappa_p V_g + V_s}{U_T}} \left(1 - e^{\frac{V_{ds}}{U_T}} \right)$	<ul style="list-style-type: none"> • Standard Subthreshold Transistor Assumption.
	Diode Connected NFET: <ul style="list-style-type: none"> • By connecting V_g to V_d, we enforce the transistor to run in saturation. • The current flowing creates a feedback loop between the drain and the charge where both automatically adapt to match each other and work saturation. • This yields that current controls gate voltage. This is very important as it is a clever way to adjust gate voltage from the current, which is typically not possible as there is infinite impedance between the channel (source, drain and well) and the gate. 	$I_{ds} \approx I_{n0} e^{\frac{\kappa_n V_g - V_s}{U_T}}$	<ul style="list-style-type: none"> • Standard Subthreshold Transistor Assumption.
	N-Type Current Mirror: <ul style="list-style-type: none"> • Creating a "Mirrored copy" of the input current. 	<ul style="list-style-type: none"> • No gain: $I_{in} = I_{out}$	

 <p>A</p> <p>Circuit diagram showing an operational amplifier (A) with an input current I_{in} and a feedback capacitor C_{fb}. The output V_{out} is connected to a reset transistor with threshold V_{thr} and current I_r. A memory capacitor C_{mem} is also connected to the output.</p> <p>B</p> <p>Voltage trace showing V_{mem} (green) increasing linearly from 0 to V_{dd} during the interval $[t_0, t_L]$, and V_{out} (red) being high during this period.</p>	<p>Axon-Hillock Circuit:</p> <p>Emulation of an integrate-and-fire neuron model</p> <ul style="list-style-type: none"> At the beginning, $I_{in}, V_{mem} = 0 \implies V_{out} = 0$ and the reset transistor is turned off ($I_r = 0$) For a constant input current I_{in}, the capacitors charge up and V_{mem} increases linearly When $V_{mem} > V_{thr} \implies V_{out} = V_{dd}$ and the reset transistor is turned on ($I_r > 0$) As $I_r \gg I_{in}$, the capacitors are discharged and V_{mem} decreases linearly until $V_{mem} < V_{thr}$ Drawback: Large power consumption if two inverting amplifiers are used as a non-inverting amplifier due to slow switching times (as proposed in original circuit) 	<p>Spike time interval ($f = \frac{1}{t_L}$)</p> $t_L = \frac{C_{fb}}{I_{in}} V_{dd}$ <p>Pulse width:</p> $t_H = \frac{C_{fb}}{I_r - I_{in}} V_{dd}$ <p>Positive Feedback</p> $\Delta V_{mem} = \frac{C_{fb}}{C_m + C_{fb}} V_{dd}$	<ul style="list-style-type: none"> Standard Subthreshold Transistor Assumption Operation in saturation regime
 <p>Conductance-based Silicon Neuron:</p> <p>Emulation of a biologically plausible conductance-based silicon neuron</p>	<p>Circuit diagram showing three components: Passive, Sodium (positive feedback), and Potassium (negative feedback). Each component consists of an inverter and resistors.</p> <ul style="list-style-type: none"> It consists of three components: passive, sodium (positive feedback) and potassium (negative feedback) Drawback: 	<ul style="list-style-type: none"> Standard Subthreshold Transistor Assumption Operation in saturation regime 	