A.A. 2012/2013

Università degli Studi di Padova

- School of Engineering -

Electrical Engineering Master's Degree

Master Thesis

implementing homeostatic plasticity in analog VLSI

And how to obtain ultra long time constants on compact silicon substrates.

Candidate:
Giovanni Rovere
1033455-IL
giovanni.rovere@gmail.com

ETH zürich
Universität Zürich UZH

ETH-UZH Supervisor:
Prof. *Giacomo Indiveri*

UniPD Supervisor:
Prof. *Andrea Gerosa*

Abstract

Neuromorphic engineering systems are electronic devices that emulates the spike based computational paradigm observed in biological neural networks. The neuromorphic computing power originates from the number of artificial neurons and from the interconnections among them. Hence high density CMOS analog VLSI are an optimal implementation medium for neuromorphic hardware systems.

However, high chip integration and CMOS processes scaling yield mismatch and non-ideality phenomena that limit the performances of the device. A neuromorphic approach to address this problem is to implement the Synaptic Homeostatic Plasticity (SHP) in silicon. SHP is a property observed in real neurons that modifies the synaptic gain in order to stabilize the neuronal firing rate activity in face of instance variations and stimuli changes.

In engineering terms, the SHP is equivalent to an Automatic Gain Control (AGC) loop comprising a Low Pass Filter (LPF). Such LPF must have a cut-off frequency several order of magnitude lower than the neurons dynamic in order to not interfere with the learning mechanisms. However, due to integration reasons, long time constants must be obtained exploiting low currents rather than increase the capacitor area. State of the art homeostatic plasticity implementations exploit floating gate devices or off-chip workstation control systems that require additional circuitry or prevent from the use in low power portable applications.

Given such LPF challenging specifications, I developed a compact CMOS filter architecture based on leakages currents in a pMOS device. I carried out and reported simulation measurements that shows AGC time constants on the order of minutes with 1pF capacitor.

A Mamma e Papà, per il loro affetto e sostegno immenso.
Alla Famiglia e agli amici più cari.
A F.A., G.M ed E.C. a cui devo veramente molto.

CONTENTS

Preface

This master thesis work was carried in the Neuromorphic Cognitive Group (NCS) under the supervision of Prof. Giacomo Indiveri of UZH|ETH at the Institute of Neuroinformatics (INI), Zürich. The research and writing work has been conducted from late February 2013 to September 2013 and is part of the Erasmus framework. The aim of this thesis report is to document the work carried out, the results I found and the whole design process in which I was involved in.

Analogue VLSI design is an iterative process that requires countless changes in the topology, sizes and architecture of the design layout in order to convey form the original idea to the application. Hence, in this thesis in addition to the final working circuit I will report also other circuits I simulated, but that showed insurmountable issues and didn't meet the specifications. However, I reported here only two attempts, the ones I believed were more meaningful due to their initial appealing. Countless other solution were subject of my research both from taken literature and from scratch.

I would like to emphasize that with this work I had the invaluable opportunity to deal with the classical analogue design work-flow (idea, design, simulation, layout). This work required considerable effort and concentration, especially in order to find a robust and reliable circuit that could be integrated in a bigger chip (MN256r1). Finally, at the end, the circuit was successfully taped out and the chip is expected to be given back on Fall 2013. Unfortunately for this reason, I can't include the real circuits measurements on this thesis report, but its test and characterization will be carried afterwards.

In addition to this main work I designed (on early March 2013) a rail-to-rail OPAMP that will be used as replacement for the old "WidePAD". It is a buffer for monitoring chip signals. It will be used in the unity gain arrangement, so true rail to rail input and output were key specifications. Hence, a two stage OPAMP is designed and integrated in the same chip (MN256r1).

This thesis report is structured as follow: Chapter one gives the context and background of neural networks and define the thesis aim. Chapter two briefly analyses meaningful neuromorphic circuits and concepts exploited later in the design. Chapter three focuses on the circuit design comparing different solutions. In chapter four is finally presented the on chip implementation of the circuit with Cadence simulations. Chapter five closes this report with conclusions and further considerations.

Except when otherwise explicitly mentioned, circuits, insights and arrangements found on Chapter three and Chapter four are, as far as I know, original results that I obtained. Due to the newness in the solutions, new concepts and circuits has been condensed in a paper that will be submitted to IEEE International Symposium on Circuits and Systems (ISCAS) 2014.

In addition to prof. Giacomo Indiveri for his invaluable support, I would also thank Qiao

Ning (INI, ETH) and Chiara Bartolozzi (IIT, Genova) that really helped me a lot, and of course the whole Institute of Neuroinformatics.

CHAPTER

1 INTRODUCTION

This chapter provides an overview of basic concepts that will be used later in the discussion. These sections aren't complete and thorough and intended only for contextualize this thesis work. This introduction ranges from the biological neuron and ends with a brief discussion about neuromorphic hardware and circuits.

1.1 Neurons

The neurons are electrical excitability cells that are specialized in intracellular communication. They are responsible for processing information over long distances through electrical and chemical mechanisms.

A typical neuron possess a cell body called the soma, several dendrites and the axon. The neuron internal volume is separated from the surrounding environment by the cellular membrane.

The **soma** is a compact section which contains the nucleus and several other organelles, which besides, provides the energy to the neuron.

The **dendrites** are thin bulges that arises from the cell body forming a complex structure and branches. These bulges are generally thin and scattered far away from the soma, hence its structure is often called dendritic tree due to its shape that resembles the roots of a tree. The dendritic tree dimensions and shapes can be very variable even within the same organism according to the neuron purpose and its location in the body.

The **axon** is another bulge that still originate from the soma but is much longer than the dendrites. Its diameter is roughly constant and it eventually ends with multiple terminations. A stylised depiction of a biological neuron is given in Figure 1.1.

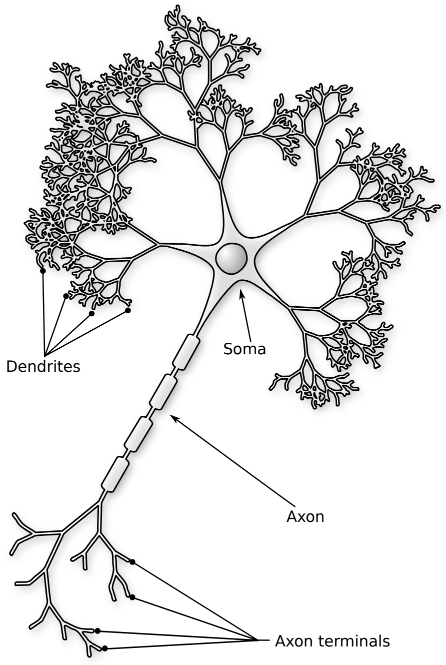

Figure 1.1: The biological neuron. Adapted from Nicolas Rougier, 2007 🅭🅯🄎

The neuron is identified as the primary functional unit of the brain and of the nervous system in general. From the functional point of view the dendrites are the neuron inputs weather the axon terminals are the output gateways of the neuron. The input signals, gathered by the dendrites, are summed, conveyed into the soma and then integrated over time by it. Next, the resulting value is compared with a threshold value. If the signal exceeds that threshold then what is called an **action potential** is established and the neuron **fires**.

An action potential is a narrow spike generated in the soma that runs through the whole axon. From the mathematical point of view a biological neuron can be model as an integrate and fire system.

To note that signals coming from the dendrites can have different weights. This is due because the path from the dendrite to the soma is passive. The strength of the signal at the soma point (before the soma integration and sum) can be computed by using the passive cable theory, whose parameter depends on the physical dimension of the path from the dendrite through the soma. A block representation with n inputs and one (y) output is given in Figure 1.2.

But how these signals are created in biological neurons? These neuronal signals are represented the voltages set up by the ions gradients across the neuronal membranes. These membranes, made by a lipid bilayer with embedded proteins, are selectively permeable to ions and separates the interior of the neurons form the outside environment. These proteins acts as active ion pumps and as ion channels. The membrane has these peculiar abilities that are responsible for most of the neuronal properties.

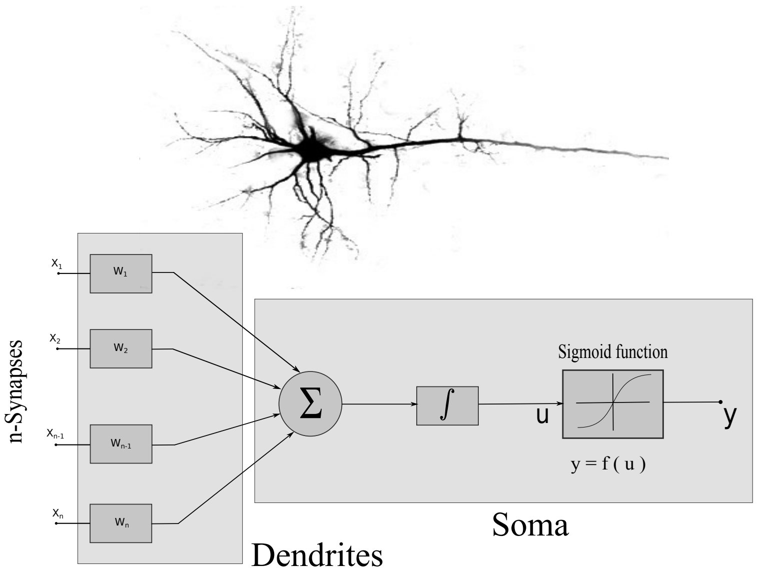

Figure 1.2: The mathematical model of the biological neuron, here depicted with n synapses and one axonal termination. u is the synpatic drive while y is the output firing rate.

	Intra	Extra
Potassium (K+)	140	5
Sodium (Na+)	10	145
Chloride (Cl-)	20	110
Calcium (Ca2+)	0.0001	2

Table 1.1: Intracellular and extracellular concentrations of ions in mammalian neurons [mM]. Taken from [19].

The **ion pumps** can actively moves ion against concentration gradients weather **ion channels** allow certain kind of ions to diffuse according to their concentration gradient. Thus, two opposite flows of ions can be identified and alter the internal ion concentration in the neuron. The external concentration is assumed to be constant due to its big volume compared to the neuron body volume. Even though this ion flow mechanism will continue forever, it will reach an equilibrium point in which the two ions flow ratios get equal. Since this equilibrium point is reached with different concentrations of ions (inside and outside the neuron), a voltage across the membrane is set.

There are mainly four ion types inside and outside the neurons, namely Sodium (Na+), potassium(K+), chloride(Cl-) and calcium (Ca2+). At the equilibrium point the intracellular concentration of potassium is grater than the other ions concentrations (see Table 1.1) and the membrane voltage potential can be calculated by the Goldman equation (1.1).

$$V_{mem} = \frac{RT}{F} ln\left(\frac{P_{Na^+}[Na^+]_{\text{out}} + P_{K^+}[K^+]_{\text{out}} + P_{Cl^-}[Cl^-]_{\text{in}}}{P_{Na^+}[Na^+]_{\text{in}} + P_{K^+}[K^+]_{\text{in}} + P_{Cl^-}[Cl^-]_{\text{out}}} \right) \tag{1.1}$$

In which P is the permeability of the membrane to a certain ion and values in brackets are the concentrations of the ions. R and F are the ideal gas constants and the Faraday's constant, while T is the temperature in Kelvin.

Hence, if computed the membrane potential with the concentration values shown in Table 1.1, the voltage result to be around $V_{mem} = -70mV$ (the positive is referred to the external of the cell).

> What happen across the cell membrane is a common mechanism in physic that has several analogies. Just for bridge it to the electronics world I give here an example on p-n junctions. P-n junction is the interface between a n-type silicon and a p-type silicon. Each of these two material have different densities of holes and electrons due to the doping processes. While in contact, a carriers gradient is established and a carriers flow from the material with higher density to the one with lower density. This motion creates a flux of carriers and meanwhile leaves ionized atoms that have an electrical charge. These atoms are static, in the sense that they can't move from their position in the silicon lattice and their charge is constant. Hence, these static ions set up an electrical gradient that produces an opposite carrier flux and counterbalance the diffusion gradient flux. At equilibrium these flows are equal and the net mobile charge is zero. However, in this condition, a voltage is measurable across the silicon interface. It can be calculate by the Boltzmann distribution, that is a special case of the Goldmann equation.

When the integrated signals come from the dendrites and summed each other, they affect the intracellular ions concentrations and thus alter the membrane potential in the soma. Hence, if a threshold voltage is reached, a spike called **action potential** is generated.

But how an action potential can be generated by the neuron? This can happen because the ion channels permeability is modulated by the ions concentrations in the neuron body. In fact, ion channels are sensitive to the membrane potential and hence their ion conductivity changes accordingly, some in an increasing way and some in a decreasing way. So, this modulation in conductivity results in a change of ion concentrations gradients and hence in a differential of potential across the neuronal membrane.

Usually there are several type of ion channel that exhibits different dynamics each other, but two of them are dominant. For instance, while the threshold is reached, Na+ channels quickly open, letting Na+ ions flow increase and thus altering the membrane potential up to a positive value about $+40mV$. Consequently, the K+ channels opens increasing their conductivity, counterbalancing the previous Na+ effect and bring back the membrane potential to its resting state.

These and other dynamics are summed over the time and, if measured by an oscilloscope, they result in a spike voltage, see Figure 1.3. In this spike we can recognize different phases, such as depolarization of the membrane, repolarization of the membrane and a refractory period that reset the neuron to the resting state. These phases are the direct result of the ion channels dynamics.

To note that these spikes amplitude and shape are invariant to the stimuli from which are generated (as far as the threshold voltage is reached). This is what is commonly referred as "all or nothing".

After being generated, the spikes runs though the whole axon and reach the end without loosing their intensity. In fact, throughout the axon, the signal is restored after fixed distances, in order to prevent signal degradation. This resemble a serial digital communication through a single line.

Actually this is a very simple model that doesn't takes into account several other dynamics. For example neurons exhibits refractory periods, it means that the interval between spikes in response of a train of stimuli gets higher and higher. Or even the homeostatic properties, subject of this thesis work, that will be explained later. However, the mechanics of all these dynamics arise from the membrane ion permeability modulation.

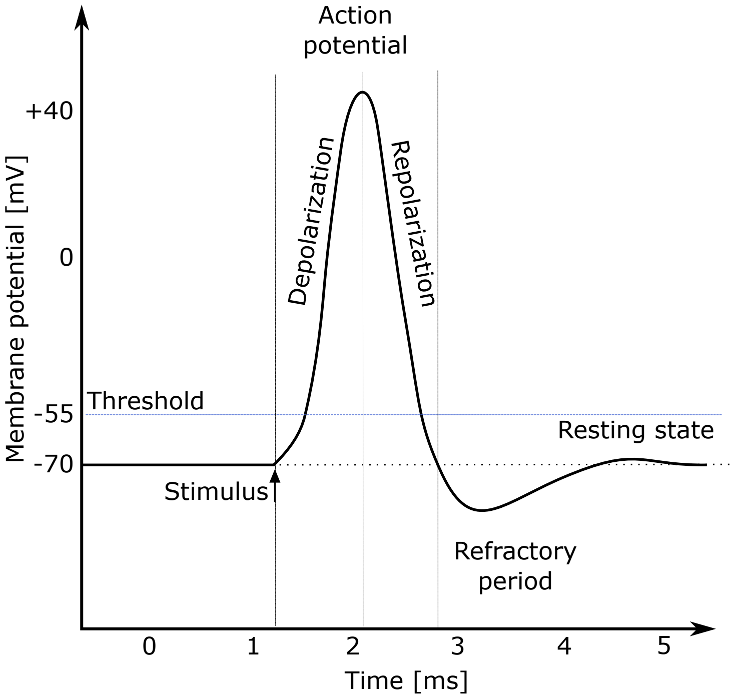

Figure 1.3: The qualitative action potentnial of a generic neuron.

Of course neurons doesn't like to stay alone and prefer to gather all together and forming a complex network. Usually neurons are shown to be connected each other in order to form a network. In this network, the neural output signals generated by the soma, runs through the whole axon up to the end of it. Then, an electrochemical connection called **synapse**, connect the terminal part of the axon to the dendrite of another neuron, establishing a connection between two neurons. This happens countless time in a complex nervous system resulting in a tangled net in which information propagates from a neuron to another one and so on.

The synapses are hence a biological structures that allow a neuron to communicate to another neuron. Synapses can be both excitatory or inhibitor. It means that they can "decide" weather the received information signal from the previous neuron axon will contribute in a positive (or negative) way making the neuron more (or less) likely to fire an action potential.

1.2 Neural Networks

An artificial neural network is a mathematical model inspired by biological neural networks that consists of interconnected groups of artificial neurons. It is based on simple process elements (neurons) and it exhibits a complex behaviour determined by the connections between the elements and their relative connection weights. Neurons in neural networks can be labelled in cluster called layer. Those layer differ each other from the functions that perform in the networks. For example the feedforward neural network shown in Figure 1.4 consist of three layers of neurons. The input layer is responsible to gather the stimuli form the external environment, the hidden layer is responsible for the process of information while the output layer provides the processed signals to the external word.

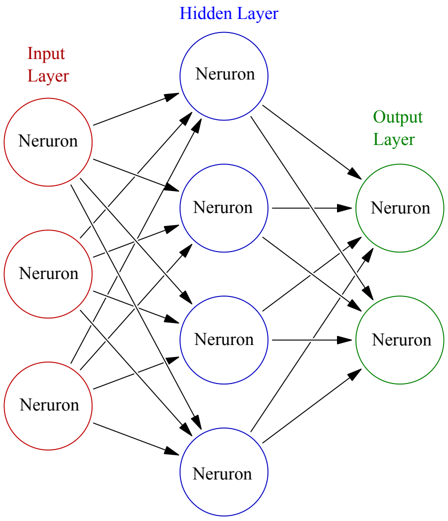

Figure 1.4: A schematic of a feedforward neural network with three layers. These layers consists of 3, 4 and 2 neurons each. Every neuron in a layer is connected to all the neurons in the previous layer.

One of the main characteristics that biological neural networks have is their **intrinsic plasticity**. A neural network is said to be plastic if the strength of the connections among neurones can be somehow modified. In the artificial neural networks model these strengths are stored in parameters called synaptic weights. These parameters can be updated according to the system input and to other events. Generally the synaptic weight can range from -1 to +1. If it's negative then the synapse is called inhibitory, if the weight is positive the synapse is excitatory (recall end of Section 1.1). These different weights process the data in the calculations and are responsible to the different behaviour of a neural network. One of the most interesting and promising application of neural network is to develop machine learning algorithms.

Learning means, given a specific task and a class of functions, to use a set of observation in order to find function in the given class that solves the problem in an optimal way. An example of learning in neural networks is the capability of these systems to recognise letters and numbers in real time, even with difficult backgrounds. This task is performed without having explicitly programmed the neural network, but simply providing to it a sufficient input statistic. Others appealing applications of neural networks ranges from the speech recognition and classification to robotic sensing and data processing.

These great capabilities of these systems doesn't lie in the neuron level but, on contrary, at system level. This power of calculus is strictly related to the number of the connections of the system and to the network plasticity properties.

The plastic system ability to update its weights is governed by this simple rule: *connections between two neuron are reinforced if they fire at the same time, otherwise weakened.*

This property is called synaptic plasticity and is first developed by Donald Hebb in 1949, thus the name Hebbian Theory. We can have a Long Term Potentiation (LTP) if their strength increases over time or a Long Term Depression (LTD) if is weakened over time. These weights are modified by the network itself according to its input statistic (unsupervised learning), and the end of this learning period the weights of the system are set and their value shouldn't vary remarkably on time. Hence, **relative weights among synapses are critical and contains information.**

Synaptic plasticity is probably the most important feature of neuronal networks. This peculiar characteristics of the neural networks is the basis of memory and their ability to learn without being explicitly programmed.

In biological neural system these connection weights depending both on the physical distance of the synapses to the soma (see Section 1.1) and on chemical proprieties of the synapses. But these weights aren't fixed, in fact the synaptic strengths have the ability to change their weight according to the received stimuli.

After this short introduction should be clear that, unlike von Neumann model computations, artificial neural networks process signals in a highly parallel way. An independent program memory doesn't really exist and process flow and memory are shared in the neural network synaptic weights.

1.3 Neuromorphic Engineering

The neuromorphic engineering is the branch of the electrical engineering that studies how to mimic neuro-biological architectures present in the nervous system and reliably build it on silicon. The original concept was developed by Carver Mead at Caltech in the late 80's and condensed in [1].

In particular, the neuromorphic engineers build circuits that emulates the same behaviour observed in biological neurons, retinas, cochleas and so on. Then these basic circuits are used as building block for making networks of these instances in order to process information. However, one of the main goals of the neuromorphic engineering is that not only these building blocks should behave as observed in nature, but even at system level this philosophy should be applied. Hence, neuromorphic circuits should be connected as close as nature arranges neuronal cells in the nervous system. This close similarity in the operation allows the engineers to build efficient systems as nature does.

But, on the other hand, neuromorphic hardware are a useful tool for neuroscientists too. In fact they can perform experiments that otherwise would be impossible to carry out with real neurons. For instance they can exactly define how many and how connect their population of neurons, perform experiments and evaluate their hypothesis. It's a kind of symbiotic interaction between engineers and neuroscientists in which both take advantage from this collaboration and this is one of the reason this emerging field is become so relevant.

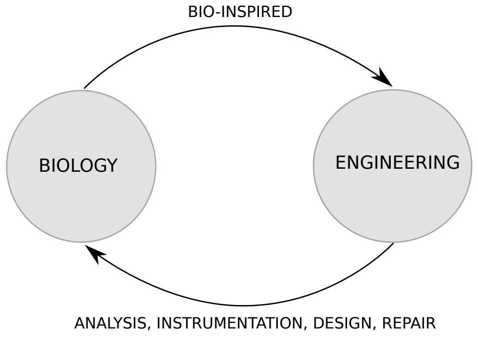

Figure 1.5: Biology and Engineering symbiotic interaction. Rahul Sarpeshkar.

Here I report what I believe is a meaningful example about how nature inspired us fixing technological issues. In addition to that, this particular example is in strict relation with the focus of my thesis work.

> In nowadays silicon processes the MOS dimensions are approaching to the minimum physical value. This means that a regular MOS has geometrical dimensions, namely W and L that are shrinking down. Currently the MOS minimum size is just some order of magnitude wider than the silicon atom. This is an advantage for the number of integrated device you can fabricate into a single die but it's a serious problem when silicon non homogeneities

> arises. In fact, imperfections such as impurities, flaws in the silicon crystal lattice and asymmetries in process layers would generally degrade circuit performances. Indeed same instances of designed elements (MOS, capacitor, resistors) would behave differently from each other due to material variance. This is the well known **mismatch** problem. These differences must be taken seriously into account by designers because usually sum each other and hence lower the performances of the whole system.
>
> Since these problems arises with minimum size transistor, one possible solution is to make the area of sensitive devices bigger. For instance, for a MOS you can times both W and L by the same amount, let's say α. This is an effective way, even though it has some drawbacks, but is not how Nature would fix the problem. In fact, also in the real world is hard to have a perfect material up to the atoms scales, and this happens in our brains too. Our neurons are far to be exactly an precise copy of each other. But Nature doesn't fix the problem by simply making bigger neurons, just because it would such a waste of volume. Neuroscientists speculate that Nature has designed our brain with countless connection between neurons also for overcome this mismatch issue.

The computational power of our brain doesn't lie at the neuron level, but on contrary must be tried to find at network level. These countless connections are the key of the wonderful brain behaviour and capabilities.

Recalling our example, neuromorphic designer should fix mismatches of silicon devices not increasing their size, but on contrary keeping the area compact and massive interconnect the neuronal instances.

Another Nature mechanism that helps to fix the mismatches problem is the network plasticity. Plasticity in the brain is not only related to the learning mechanisms, but even modifies the strength of neurons in order to keep the neuronal network in working conditions in face of system variations, such as mismatch. Implementing this plasticity in hardware would hence be another great way to apply **bioinspired mechanisms** in silicon design. This type of plasticity is called homeostatic plasticity and will be further discussed in section 1.4.

However, the main reason that urged early scientists into this insight of the neuromorphic engineering is that, even with the most powerful supercomputer available, we can't perform tasks such as pattern recognition, classification, motor control and so on **as efficient and as effective as animals do**. The reason of this huge difference is that computer and brains are based and relied on different principles and paradigms.

In the next section I will first highlight these differences between brain and non neuromorphic computer (supercomputer), then I will give a quick overview about different approaches and philosophies that drives neuromorphic engineers. The further discussion about neuromorphic hardware will be focused especially on the neuromorphic cognitive systems rather than neuromorphic sensors or interfaces. I.e. architectures that can reason about the actions to take in response to the combinations of external stimuli, internal states, and behavioural objectives. These hardware are directly inspired by the brain.

1.3.1 Supercomputer vs neuromorphic hardware

As already shown in the previous section the brain processes the information mainly in an analogue fashion. The key entity that carries information are the spikes and those signals are processed throughout a 3D (volume) neural network.

As contrary, supercomputers information are always associated to digital values and their spatial domain is a flat piece of silicon, reducing the networking capabilities. These two systems, the neural networks and the supercomputers, relies on completely different architecture. The supercomputers are based on the von Neumann paradigm, i.e. there is a digital CPU that process input data according to the stored program located somewhere next the CPU. This architecture is hard-wired and modifications in the topology aren't usually expected. The instructions fetch and execute flow is heavily sequential and no more than one operation can be performed at once (with one core). However, even though these machine could be theoretically made in the analogue domain, their natural implementation is the digital clocked domain. Digital is because the information are associated by a sequence of zeros and ones and clocked is because the timing of the operation is precisely set by a high frequency reference (around GHz in modern devices). The strength of this paradigm is that it's really robust to noise and can easily handle large amount of data. But is very power hungry and for some particular tasks its very inefficient. Usually, supecomputer are excellent in providing exact solution to well defined problems.

Neuromorphic hardware are completely different. They are based on analogue devices and signals are sort of digitalized only for transmit information inside and outside the chip. Digital signals are not involved in the direct computational process. The architecture is based on neural networks that are massively parallel, it means that all the artificial neurons works together simultaneously. The computing timing is not set any more by a clock, but is performed in an event-driven manner.

Another big difference is that in supercomputer you have to write a very precise and flawless code in order to get proper behaviour. On neuromoprhic cognitive hardware you don't have a proper program but you use the learning paradigm to allow it perform useful tasks.

Another great result about how the Nature build brains is that they can still work even if a neuron is corrupted, supercomputer could became useless even if a single transistor gets out of order. This nice characteristic can be observed even in large artificial neural networks and is really useful in order to face unavoidable fabrication failures or inhomogeneities.

From this discussion is clear that we can model pretty well the dynamics of a single neuron. Thus, simulate a neural network by setting equation of a supercomputer would result in exact results and on the same behaviour of the modelled neural network. But the power consumption and the computation time would be exaggerate, prevent it from the use of real time and useful computation. On contrary, neuromorphic hardware **emulates** rather than **simulate**.

Weak Inversion	Vs	Strong Inversion
Voltage-mode	Vs	Current-mode
Non-clocked	Vs	Switched Capacitor
Biophysical model	Vs	Phenomenological model
Real-time	Vs	Accelerated-time

Table 1.2: Silicon neurons design styles.

Emulation and simulation are two different concepts. The simulation is a conceptual model that represents the reality and only reproduce a certain input-output black block behaviour. An ordinary supercomputer can simulate neuronal networks in the sense that it behaves like that, even though the underlying mechanisms are totally different. On the contrary the emulation aims to exact mimic a system behaviour in a lower level, even though it's based on different means.

1.3.2 Silicon Neurons

Several artificial neurons have been developed since the very beginning. Of course, as on every branch of engineering, trade offs on specifications are unavoidable, hence different design styles were developed. The following comparison is extracted from [2] and summarize the mains neuromorphic techniques to build silicon neurons.

From the structural point of view silicon neurons can all be described by the schematic of Figure 1.2. The synapses receives the spikes, integrated them over the time and sum each other output currents. The soma block get the summed current from the synapses, and if this current crosses a threshold, an event, namely a spike, is generated at the output of the neuron.

As already emphasized, the neurons exhibits a complex dynamics over time. To mimic such complex behaviour an artificial neuron requires more circuitry and a relatively big area. This reduces the possibility to pack several neurons into a chip and then simulate large neural networks. Hence, several neuronal model were developed, some of them are cumbersome but behave very realistic, some other tends to be small but have less faithful dynamics.

Some opposite techniques that are common among neuromorphic engineers are recap in Table 1.2. Each of the computational blocks of Figure 1.2 can be implemented in any of these design styles.

The **weak** and **strong inversion** are referred to the operational point of the MOS transistors. Strong inversion is the usual above threshold bias point. Weak inversion is also called sub-threshold. In this condition the channel is not completely formed . Usually the strong inversion requires more current to bias the MOS, is much more insensitive to mismatches and the I_{ds} vs V_{ds} is a square law. The weak inversion is more suitable for low power operation and its I_{ds} vs V_{ds} is an exponential curve.

Voltage mode or **current mode** circuits means that the meaningful variable is represented respectively by a difference in potential or in a flow of electrons. Current mode circuits are more appropriate for low power operation because are more insensitive to power supply reduction trend.

Non clocked signal are continuous in time, conversely processing could be done in analogue domain but in quantized **clocked** time . For instance this happens in switched capacitor circuit. These circuits work with a clock with two phases. In the first phase, charges that carries information are stored in a capacitor and (ideally) can't move. Then, in the other phase, switch are closed and the charges can flow freely according to the circuit dynamics set by the topology. This technique is more complex but allow better matching and precision.

Biophysical and **phenomenological** models refers to the level of the detail on which circuits emulates neuronal behaviour. As already pointed out, this trade off is strictly related to the used area and then to the number of neuron you can integrate on a single die.

Circuits that can operate with time scales that are biologically comparable (on the order of magnitude of milliseconds) are named **real-time**. Conversely circuits that runs with time scales, at least, 10 times faster are said to be at **accelerated-time**. This distinction is pretty important, in fact, even though simulation of neural behaviour could be run event at accelerated time, the interaction with real world through sensors should be efficiently performed only in real-time scales.

An example of silicon neuron is the one shown in Figure 1.6. This is the conceptual schematic of neurons that are integrated in the MN256r1 chip developed by the NCS group.

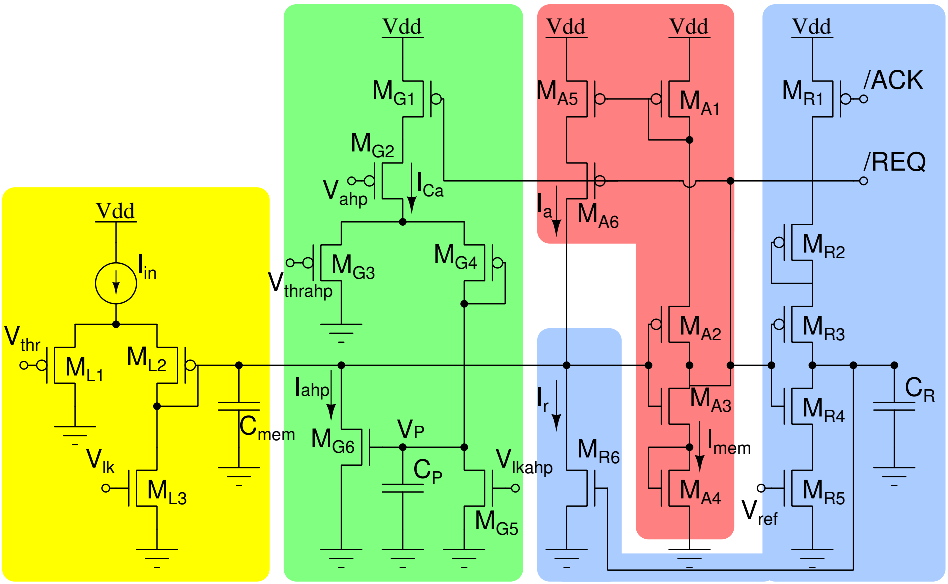

Figure 1.6: An integrate & fire artifial neuron based on the DPI [5].

This artificial neuron consist of four blocks. The yellow block is the synapse of the neuron and is based on the Differential Pair Integrator circuit (a detailed explanation of this circuit will be provided in section 2.4). This DPI takes the current input, integrates it and charges the capacitor C_{mem} accordingly. The DPI acts as the neuronal input synapse while the voltage across C_{mem} is the membrane potential. To note that a single neuron can have several input synapses, hence several DPIs in parallel.

The red part is the spike event generation block. The blue block is in charge for the reset of the neuron and for providing digital acknowledgement and request signals for the inter chip neuronal communication (AER). The green block is another DPI (in this case it doesn't act as a synapse) in charge of set the temporal dynamics of the neuron.

According to Table 1.2, this circuit can be categorized as a weak inversion, current mode, non-clocked, phenomenological circuit. Real or accelerated time can be choose according to the capacitor values.

1.3.3 MN256r1 chip

The MN256r1 chip is the chip developed by the NCS group at INI. Its main purpose it to emulate a medium size neuronal network with plastic and cognitive capabilities.

It consists of 256 neurons, each of which has 256 plastic synapses (DPIs) and 256 non plastic synapses (basically is a simpler version of a DPI in which the weights are not settable). In order to form a neural network all these neurons are connected each other. This is a challenging task due to large amount of connections in the chip. This routing is implemented via a the AER protocol(Address-Event-Representation). Basically it's a sort of "virtual wiring" technique for interconnecting spiking circuits in neuromorphic systems. In the AER protocol, whenever a spiking neuron in the chip generates a spike, its "address" is written on a high speed digital bus and sent to the receiving neuron(s). To note that here the digital part doesn't interfere with the signal processing.

The chip has a 64 programmable bias generator, each of those signals are independent and whose current ranges from $50fA$ to $20\mu A$. The chip die size is 9mm x 5mm.

In order to crate a cognitive neural network a full understanding about real neurons at cellular level and their complex connections are required. But, even technological aspects are important and were one of the bottle neck that delayed the development of neuromorphic hardware. However, is clear that this new emerging field of science is a beautiful merge of different topics and requires a convergence of knowledges in order to go further.

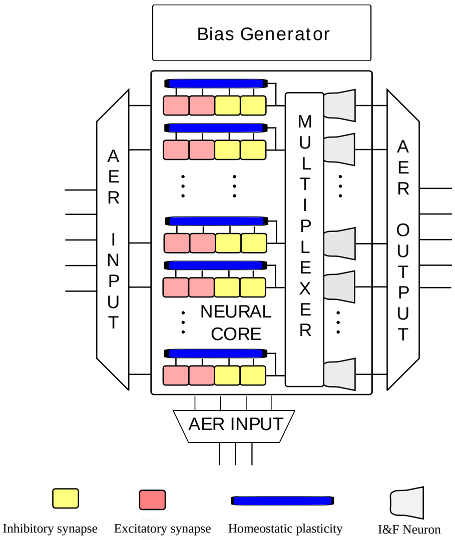

Figure 1.7: High level conceptual schematic of the MN256r1 chip.

1.4 The Homeostatic Principle

The "Homeostasis" word originates from the union of two Greek words: hómoios = similar and stásis = standing still. Homeostasis is a property of a systems that tends to hold its own behaviour in a stable condition in face of environmental fluctuations or input changes. It's a different concept from the *"dynamic equilibrium,"* which is not regulated and from *"steady states,"* that are stable but can jump into different states if energy is applied. Indeed, homeostasis forces the system to its only equilibrium point which is asymptomatically stable.

The homeostatic principle is a general mechanism of physiology that was firstly observed by Cannon in the late 20s. Way more recently, Turrigiano et al. [4] showed that this also applies to a population of plastic neurons.

In [4] is showed that this principle is fundamental for maintain the stability of a neural network and let the whole system works properly. As already emphasized, the leaning mechanism of neural networks based on the Hebbian principle updates the synaptic weights. It is basically a positive feedback, in the sense that it tends to destabilize the system forcing the synaptic weights to reach their maximum or minimum value (±1). Electronically speaking it means that the synapses approach respectively a high or low gain. Hence, these mechanisms of plasticity will lead the system towards runaway excitation or quiescence.

One way to visualize this problem is consider a neural network made by five layers connected in a feedforward way as show in Figure 1.8. There you can see that if the gain between two adjacent layers is too high, then with a certain input, the network output will saturate and lost its information. It means that the firing rate of the last layer neurons will be always somehow at high firing rate and almost independent from the input signals. Vice versa, if the gain is too low the system will be pushed into an high attenuation state and the output firing rate will approach to zero over time.

This problem directly arises because the synapses are highly plastic and their weights can remarkably changes according to their input signals. Hence setting initial condition in the right state would not prevent the system to fall into saturation.

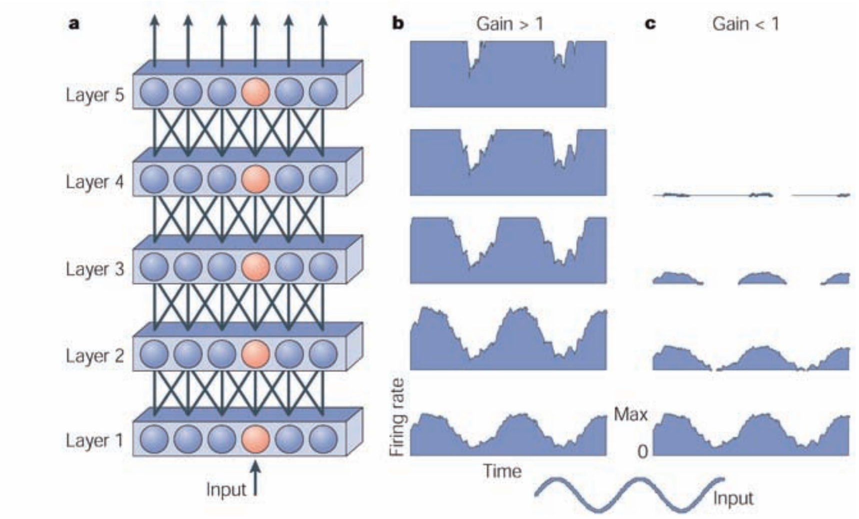

Figure 1.8: This illustrates how a gain higher or lower than 1 can distort the input signals if processed by a chain of neurons. **(a)** Neural network **(b)** Output firing rate from each layer. Saturation. **(c)** Output firing rate from each layer. Attenuation. Image taken from [4].

On contrary the homeostatic principle is a negative feedback that tends to stabilize the system in face of input changes or in environmental variations. It senses the output of the neuron and changes their synaptic gain accordingly. This principle is represented in Figure 1.9.

If we interpret the plot at a fixed time it describes the steady state (static) behaviour of the system. If the synaptic drive is low then the firing rate is low, and vice versa. The static dependence of these two variable is not linear and when the curve gets flat then information distortion occurs. This happens when the synaptic gains are really high or really low, that is the case which the Hebbian mechanism would pull towards.

On contrary, the homeostatic principle forces the synaptic drive to lie in the inner region of the plot, there the behaviour is linear. The highlighted zone is the target firing rate, so the two arrows pointing each other shows the dynamics of the homoeostatic principle.

Even though there could be several types of homeostatic principle (additive, multiplicative) the multiplicative is the one I will consider from now on. The multiplicative homeostatic principle means that it changes the synaptic drives multiplying or dividing all the synapses weights of one neuron. Hence, the relative synaptic strengths are keep constant, but the absolute strengths are increased or decreased. This is really important, otherwise the "program", learnt by the network and stored in the synaptic weights, would be distorted or lost.

Trying to draw again a parallelism to the electronics world, it's clear there is an evident analogy between the saturation of Figure 1.8 and analogue electronics filters. In fact, high

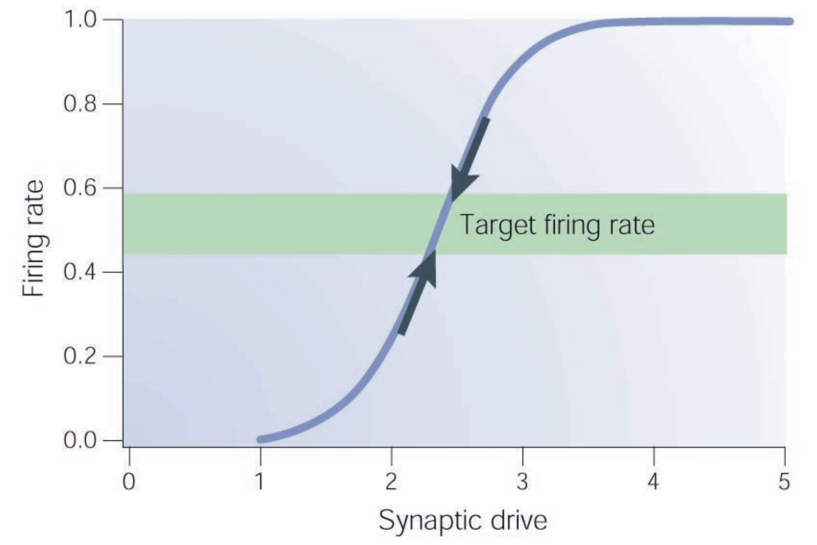

Figure 1.9: The synaptic input vs firing rate of a neuron plot. Image taken from [4].

order analog filters, consist of a cascade of second order cells. The first designer concern is the position of the poles, but it's not sufficient in order to design a reliable filter. In fact, the gain of each of those cells must be set as close to the unity as possible in order to prevent distortion and assure high a SNR. The conceptual problem is exactly the same explained before and relates to attenuation and saturation of stages.

However, analogue filter order is seldom higher than 7-8, hence the precision in the gain has relaxed boundaries. Hence, the analogue designer doesn't need to build a gain control circuit. On contrary, in neural networks the cascade of neurons is remarkably longer and this gain problems is relevant.

In addition, the homeostatic principle is not only fundamental for hold the system in a stable condition, but is very useful even facing environmental changes or imperfections (see 1.3 about mismatch).

For instance, in brain this could happen when the temperature increases (for example when you have a flu), or when ion concentrations changes due to illness or whatever else. In silicon, for example, the stability can be affected by mismatch or by variations in references (power supply variations due to plug/unplug loads in the chip). Hence, if these changes aren't excessive, the homeostatic principle would fix it.

In summary, the homeostatic principle is a global negative feedback that tends to let the system stable, it can be seen as an **Automatic Gain Control**. It has basically two benefits: it allow the system to be stable and it let the system works even in real condition (mismatches and changes in the environment).

1.4.1 The state-of-the-art

As far as I know only few attempts to implement this homeostatic principle have been made [5], [11] and [12]. In the first article the usual approach is to implement the homeostatic principle via a software that controls the hardware on the chip. This methods allows to get time constants arbitrarily long (10^4) and is useful to perform generic experiments on homeostasis in order to validate the model. Of course, these setup have limited capabilities if the final goal is to develop a very compact low power neuromorphic hardware.

The authors in [11] actually proposed an "on chip" solution. But the performances of the homeostatic loop (a few seconds) aren't even comparable to those obtained with HW/SW mixed setup of [5]. To get long time constants is crucial in order not to interfere with the learning mechanisms and to exploit neural networks behaviour. A further analysis how their circuits can't get high performances will be given in Section 3.3. In [12] the authors implemented the homeosatic plasticity on silicon using the floating gate MOS technique. Although the performance of this implementation are really high (time constants of minutes), dealing with floating gate require special processes and more circuitry.

As already emphasized, the challenging task is to get really long time constants on silicon. Actually it's difficult to achieve because, in standard CMOS process, leakage currents and second order non-idealities must be taken into account. Hence, modified topologies and structure must be examined. In addition to that, a careful layout must be drown in order to guarantee high performances.

1.5 The Thesis Aim

The aim of this thesis work is to report the design of a circuit that implements the homeostatic principle on silicon. The circuit will be integrated in the MN256r1 chip and fabricated in standard CMOS AMS 0.18μm process.

In the above brief introduction I tried to give insights focusing on trade offs that are dominant in neuromorphic engineering. Here below I report the given circuit specification that my design has to accomplish.

Desired Circuit Specifications:

- **Ultra long time constant**
  With a capacitor value of 1pF. However is not important the exact value of the time constants as long as they are at least several seconds. Variance of the time constants across chip instances is not critical.

- **Small silicon area**
  The maximum circuit available area, capacitor area excluded, is 16.5μm × 50μm.

- **Low power consumption**
  No explicit upper boundaries were give but recall that this circuit must be instantiated for each neuron (currently 256). So, the lower, the better.

CHAPTER

2

CONCEPT AND CIRCUITS OF NEUROMORPHIC ENGINEERING

The purpose of this chapter is to recall and provide the foundations of engineering concepts that will be essential in order to fully understand the design and phenomena on which my circuit is based on. A more detailed and comprehensive analysis of these topics are covered by several books and scientific papers, some of them are listed in the bibliography. Here I start with a review of the MOS device and the semiconductor physics, that is the key element of modern integrated circuits. A special operation point of this device is analysed and will be the key of my design low current generator. Then the log-domain filters technique is presented along with few meaningful examples and circuits. Many derivations are given only for nMOS type transistors, pMOS extension is straightforward.

2.1 The Subthreshold MOS

The MOSFET is an unipolar electrical active device made out of silicon that can be used both as a digital switch or as an analogue amplifier. MOSFET is the acronym of Metal-Oxide-Semiconductor-Field-Effect-Transistor that recall its physical structure. Usually MOSFETs are divided into two groups according to the type of currents, i.e. nMOS if currents consist of electrons or pMOS if currents consist of holes.

The MOS consist of different layers of doped silicon. However, the starting point is the **intrinsic silicon**, i.e. the pure form of silicon. Silicon is a semiconductor material and compared to metals has low conductivity due to low amount of free carriers, namely electrons (e) and holes (h). The concentration of these carriers ($n_i(e)$ and $n_i(h)$) in the intrinsic silicon is proportional to the silicon energy gap (E_g) and to the absolute temperature (T) according to the following law:

$$n_i(e) = n_i(h) = n_i = CkT^{\frac{2}{3}}e^{\frac{-Eg}{2k_bT}} \tag{2.1}$$

where C is a material dependent constant and k_b is the Boltzmann constant.

To dope silicon means to add donor/acceptor atoms (usually phosphorous or boron) in the intrinsic silicon lattice. Hence, the silicon is not pure any more and it's then called **extrinsic silicon** since electrical properties are modified. In fact, these two added atoms have respectively one more and one less electron in their conduction band if compared to the silicon atom. Hence a n-type silicon, doped with boron atoms, has additional free electrons. On contrary, a p-type silicon is a silicon lattice doped with phosphorus, and has more holes than electrons. Both n and p-type silicon shows an increase in silicon conductivity proportional to the number of added atoms.

The majority carriers concentration in an extrinsic silicon is known, i.e. the number of added impurity atoms, but the minority carriers follow the mass action rule:

$$n_e(e) \cdot n_e(h) = n_i^2 \tag{2.2}$$

But how is the physical structure of a MOS? Let's think then about a nMOS.

The nMOS is a planar device made on a layer of p-type silicon bulk(B) in which, under the surface, there are implanted two small regions of n-type silicon, namely the Drain(D) and the Source(S). The section between these two terminals is covered by silicon oxide (insulator) and then by polysilicon (a not so good conductor) in order to make the Gate(G) terminal. See Figure 2.1. Due to its symmetry the source and the drain terminals are exchangeable according to their potential and aren't physically located. In fact, in a nMOS device, the source terminal is the one with lower potential, while the drain has the higher one (compared to the source). Hence, the current will always flow from the drain to the source in a nMOS. Vice versa in the pMOS devices currents flow from the source to the drain.

I will start here with an insight how the nMOS works, then equations are derived later only for the subthreshold region.

For the first time let's think about a nMOS without the source and drain terminals. The resulting structure is just a three layer structure that consist of a polysilicon gate terminal, an insulator layer and then the bulk layer (p-type silicon). Actually this structure acts as a parallel-plate capacitor, in fact the gate and the bulk are the two capacitor plates, while the oxide is the dielectric.

Let's now apply a positive voltage between the gate and bulk terminals. Then, positive charges will accumulate on the gate terminal. Since same type charges repeals each others, most of the majority carriers in the p-type silicon (holes) are repelled and pulled away from the bulk-dioxide interface beneath the gate. While these majority charges can freely move in the silicon, the atoms from which they belongs are fixed in the lattice structure. Thus, these atoms became negative charged ions since they lost one of their valence band electrons.

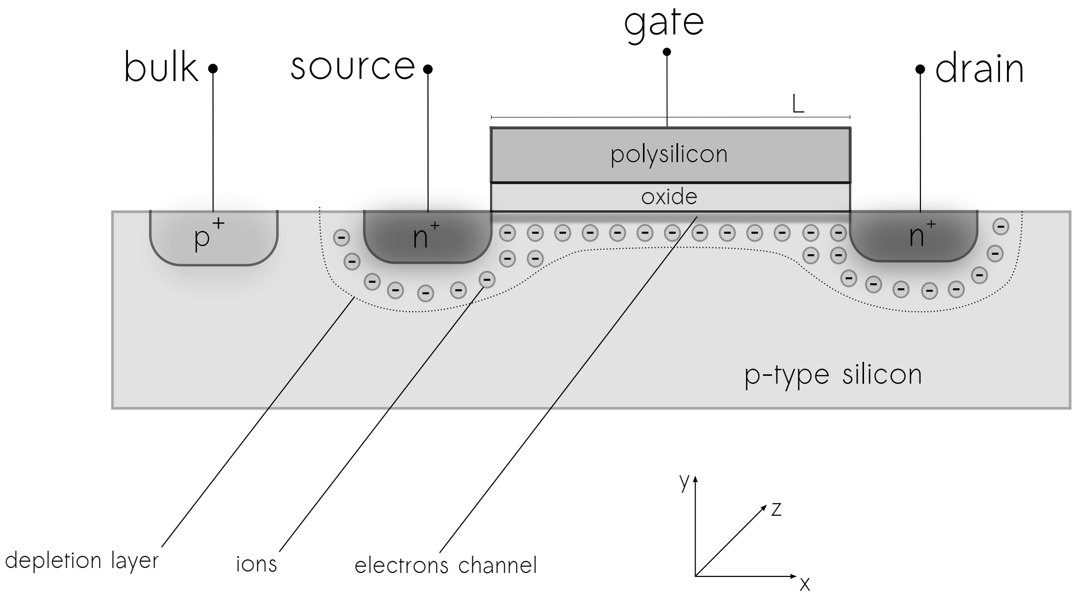

Figure 2.1: Lateral section of the nMOS in strong inversion. You can see the depletion layer and a tiny electrons channel benheat the oxide. Distinction between source and drain terminals is arbitrary since no potentials are applyed.

These negative ions counterbalance the positive charges on the gate reaching an equilibrium and forming such an ions layer called **depletion layer** (depleted by majority carriers).

At the very beginning the depth of this depletion layer is linearly modulated by the gate voltage, in the sense that the bigger is V_{gb} the bigger is the depletion layers depth $L_{depletion}$. But, when this difference in potential (V_{gs}) became so strong, even the minority carriers of p-type silicon (electrons) are attracted by the gate and, since they are free to move, they will pack and thicken just under the oxide layer. If V_{gs} goes beyond a certain threshold voltage V_{th} then the number of these minority free holes became considerable, and they start act as an additional plate that shield the depleted layer from the magnetic field created by the gate. Hence, after goes beyond a voltage threshold, the depletion layer would not increase any more, while the minority layer will.

The threshold voltage V_{th} is set by the foundry according to its own process. It's usually scaled down with the power supply scaling trend in order let MOS be able to switch on and off. For instance, in AMS $0.18\mu m$ process the regular V_{th} for a nMOS is around 300 mV. This threshold can be increased up to 550 mV using an High Threshold Voltage(HVT) nMOS. This is consist of a regular nMOS with an additional mask in the process that adjust ion implants in order to modify the threshold voltage.

Even though the threshold voltage is not adjustable by the IC designer, still it's not constant and exhibits second order dependencies. The most important effect is the so called

body effect. It consist of a modulation of the threshold according to the source-bulk voltage. This important relation is given by the following equation:

$$V_{th} = V_{th0} + \gamma \left(\sqrt{|V_{sb} + 2\phi_F|} - \sqrt{|2\phi_F|} \right) \tag{2.3}$$

where V_{th0} is the threshold voltage at $V_{sb} = 0$, and γ and ϕ_F are a process dependent parameters.

Now let's add back the source and drain terminals, resulting the usual MOS device. The before mentioned layer of minority carriers under the gate is called **channel** and effectively connect the drain and the source terminals. Then, if this channel is well established (i.e. if $V_{gs} > V_{th}$) a flow of carriers, mainly driven by drift, can rush from the drain to the source. This current intensity on first approximation depends on V_{ds}.

On contrary, if this channel is not yet formed (that means $V_{gs} < V_{th}$), still a small flow of carriers can appear from the drain to the source, but now the flow of current is mainly driven by diffusion since there is a gradient in electron concentrations between the drain-source terminals.

Now let's make clear what exactly diffusion and drifts mean.

Single carriers (electrons or holes) in silicon lattice have a complex dynamics mainly due to thermal motion and modified by multiple collisions with atoms and atomic forces. These trajectories are very difficult to analyse, but can be seen as a certain realization of stochastic process called Brownian Motion. Fortunately, currents in electronic devices still involve fluxes of high quantities of these particles and hence only their statistical average trajectories are needed in order to describe the device behaviour. Therefore, the meaningful dimensions in the MOS analysis are current densities rather than single carriers motion.

The MOS behaviour is basically governed by two kind of carrier motions in silicon. Actually they are both the results of the Brownian motion but they are studied separately [1].

The **drift motion** happens when external forces, such as an electric field (E), are applied to free charges in silicon. Hence, carriers velocity will be defined by the following formula:

$$v_{drift} = \mu E \tag{2.4}$$

in which the μ term is the mobility of the particle in a certain doped silicon lattice.

The **diffusion motion** generally arises when there is a concentration $(n(\cdot))$ gradient in the mean $\frac{dn(\cdot)}{dx}$. Particles flow from the high density region to the low density region. The diffusion velocity can be written in terms of gradient of the particle distribution along the x axis:

$$v_{diff} = -\frac{1}{2n(\cdot)} \frac{dn(\cdot)}{dx} t_f v_T^2 \tag{2.5}$$

In which equation the negative sign arises because the net flow of particles is from higher density to the lower density, $n(\cdot)$ is the particle concentration, t_f is the mean time between particle collisions and v_T is velocity average. Recalling that at thermal equilibrium a particle with mass m has kinetic energy defined by:

$$\frac{1}{2}mv_T^2 = \frac{1}{2}kT \tag{2.6}$$

hence, substituting Equation 2.6 in Equation 2.5 yields:

$$v_{diff} = -\frac{1}{2n(\cdot)}\frac{dn(\cdot)}{dx}kT\frac{t_f}{m} = -D\frac{1}{n(\cdot)}\frac{dn(\cdot)}{dx} \tag{2.7}$$

in which $D = kTt_f/2m$ is called the **diffusion constant.**

Since drift and diffusions mechanisms are different manifestation of the same phenomena, then they are related by the Einstein's relation:

$$D = \frac{kTt_f}{2m} = \frac{kT}{q}\mu \tag{2.8}$$

From these velocities, current densities through a surface $A_{surface}$ can be derived multiplying by the particle densities according to:

$$J_{diff} = n(\cdot)v_{diffusion} \tag{2.9a}$$

$$J_{drift} = n(\cdot)v_{drift} \tag{2.9b}$$

and then get currents by:

$$I_{diff} = J_{diff} \cdot A_{surface} \tag{2.10a}$$

$$I_{drift} = J_{drift} \cdot A_{surface} \tag{2.10b}$$

Subthreshold Region

As already mentioned, currents in subthreshold MOS are dominated by diffusion currents. Let's show and calculate it. The p-type silicon surface beneath the gate has potential ψ_s that is affected by the voltage applied to the gate terminal. If the gate voltage increase but stays below the threshold voltage V_{th}, no electrons channel shield effect happens and the depletion thickness $L_{depletion}$ increases proportionally in order to balance out the charges of the gate. Assuming small variation around the operating point, is true that [18]:

$$\psi_s = \psi_0 + \kappa V_g \tag{2.11}$$

where ψ_0 is just an arbitrary reference for the potential. Furthermore, κ is the gate coupling coefficient defined as $\frac{C_{oxide}}{C_{oxide}+C_{depletion}}$ and represent the coupling of the gate to the surface potential. Since C_{oxide} is fixed by the physical dimensions and the depletion capacitance

$C_{depletion}$ is fairly constant in subthreshold region, the κ is hence considered constant and typical values in modern processes are around $\kappa \approx 0.7$.

Now additionally assume that the potential of the channel, denoted with ψ_s, is constant along the whole x axis [15]. No difference in potential results no drift currents. But, since drain and source terminals are accessible from the external through physical connections, their potential can be imposed. If these potential V_s and V_d are different from V_g hence, source and drain electrons face a potential barrier respect to the gate. The barrier hight is then given by:

$$\begin{aligned} \theta_s &= \theta_0 - q(\psi_s - V_s) \\ \theta_d &= \theta_0 - q(\psi_s - V_d) \end{aligned} \tag{2.12}$$

in which $\theta_0 = qV_{bi}$ is the built in energy barrier of the two p-n junctions and q is the proton charge. These barriers are proportional to the difference between the terminal (S and D) applied voltages and the channel one.

Given the energy barrier, Equation 2.12 seen by the electrons we can compute the electron concentrations at the source and drain terminals via the following equations [1]:

$$\begin{aligned} n(e)_{source} &= n(0)e^{-\frac{\theta_s}{kT}} = n(0)e^{\frac{-(\theta_0 - q(\psi_s - Vs))}{kT}} \\ n(e)_{drain} &= n(0)e^{-\frac{\theta_d}{kT}} = n(0)e^{\frac{-(\theta_0 - q(\psi_s - Vd))}{kT}} \end{aligned} \tag{2.13}$$

Where $n(0)$ is the electron density at the reference potential.

Since the drain terminal, compared to the source one, by definition has the highest potential, hence its electrons concentration is grater than the source electron concentration $(n(e)_{drain} < n(e)_{source})$.

As emphasized over and over again, whenever there is a difference of concentration there is a gradient and hence an electron flow by diffusion, in this case from the source to the drain.

Since we know the concentrations at the ends of the transistor and we know how do they move by diffusion, hence we can compute the channel current in the MOS recalling Equations 2.10, 2.9 and 2.7.

$$I = J_{n,diff} W L_{depth} = n(e) v_{diffusion} W L_{depth} = -qWL_{depth} D_n \frac{dn(e)}{dx} \tag{2.14}$$

where $J_{n,diff}$ is the diffusion current density, W is the usual width of the channel and L_{depth} is the depth of the inversion layer, D_n is the diffusion coefficient of the electrons and $\frac{dn(\cdot)}{dx}$ is the concentration gradient across the MOS.

Since we are analysing only this phenomena, we implicitly assumes that non other currents flow in the bulk or in the gate. This is a reasonable assumption since the gate is isolated from the bulk and the drain-bulk and source-bulk junctions are usually reverse biased.

Hence, from the KLC, the current throughout the channel must be constant.

$$\frac{dn(e)}{dx} = \frac{n(e)_{drain} - n(e)_{source}}{L} = \frac{n(e)_1}{L} e^{\frac{\psi_s}{U_T}} (e^{-\frac{V_d}{U_T}} - e^{-\frac{V_s}{U_T}}) \tag{2.15}$$

$$I_{ds} = -q\frac{W}{L}L_{depth}D_n n(e)_1 e^{\frac{\psi_s}{U_T}}(e^{-\frac{V_d}{U_T}} - e^{-\frac{V_s}{U_T}}) = I_0\frac{W}{L}e^{\frac{\psi_s}{U_T}}(e^{-\frac{V_d}{U_T}} - e^{-\frac{V_s}{U_T}}) \tag{2.16}$$

where $n(e)_1 = e^{-\frac{\theta_0}{kT}}$ and $I_0 = qL_{depth}D_n n(e)_1$.
Finally, Equation 2.16 yields:

$$I_{ds} = I_0\frac{W}{L}e^{\frac{\kappa V_g}{U_t}(e^{-\frac{V_d}{U_T}} - e^{-\frac{V_s}{U_T}})} \tag{2.17}$$

As holds in the usual above threshold, even in the subthreshold regime the nMOS exhibits different behaviours according to the nMOS polarization point:

- **Non-saturation region**
  In this regions arises when $V_{ds} < U_T$ hence we can rewrite Equation 2.17 as follows:

  $$I_{ds} = I_0\frac{W}{L}e^{\frac{(\kappa V_g - V_s)}{U_T}}(1 - e^{-\frac{V_{ds}}{U_T}}) \tag{2.18}$$

  As contrary on what happens in strong inversion, here the V_{ds} vs I_{ds} curves are not linear.

- **Saturation Region**
  If $V_{ds} > 4U_T \approx 100mV$ the concentrations of electrons at the source terminal is negligible compared to the one at the drain terminal, the Equation 2.17 can be approximated with the following one:

  $$I_{ds} = I_0\frac{W}{L}e^{\frac{(\kappa V_g - V_s)}{U_T}} \tag{2.19}$$

  Intuitively, this I_{ds} vs V_d independence is consistent with the intuitive analysis. In fact, if V_{ds} is considerable, then the difference in electrons concentrations at source and terminal is big. Hence, doesn't really matters how small is the drain concentrations in the I_{ds} since their contribute to it is negligible.

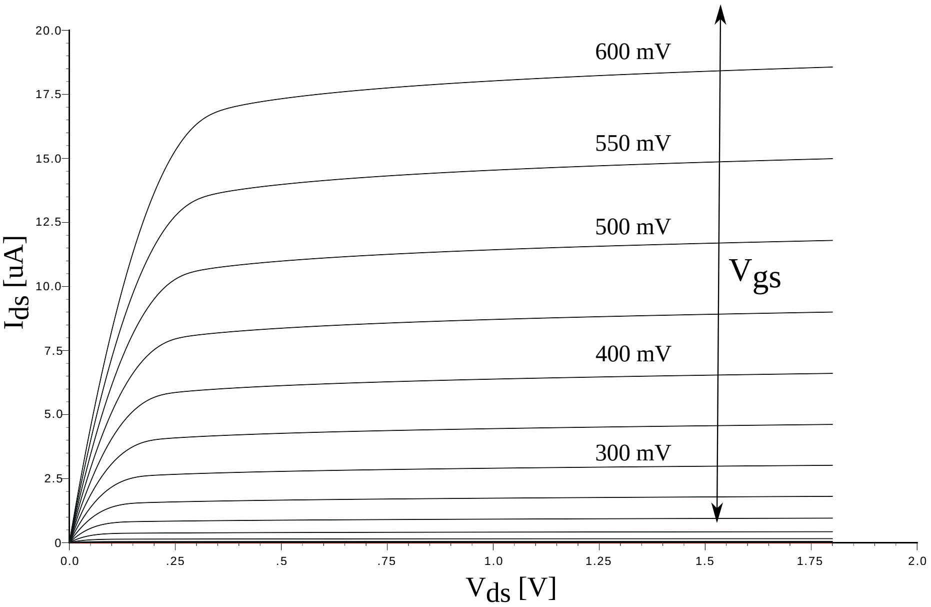

Figure 2.2: Simulation of a nMOS $\frac{W}{L} = \frac{1\mu m}{1\mu m}$ I_{ds} vs V_{ds} plot, parametrized on V_{gs} up to $600mV$ ($50mV$ steps).

2.2 A Low Leakage Cell

This section briefly summarizes and explains concepts and results reported by M. O'Halloran and R. Sarpeshkar in [6] and in [7]. Their papers focus on how to obtain accurate analogue storage cells with ultra low leakage currents. Natural applications of their results comprise analogue memories, sample and hold, switch capacitor circuits, offset cancellation and so on. However, even if they didn't mention it, their novel insight and characterization about additional leakage mechanisms in MOS devices can be used as basis for really low currents generator, exploiting parasitic leakage mechanisms in the device. As will be explained in Section 3.2, this insight is the key and technique used in this thesis work in order to get ultra low time constants.

In their papers, the authors first analysed past low leakage cells and made a fair comparison between the designs, normalizing the meaningful quantities. Later, they analysed accumulation mode MOS with in silicon measurements in order to develop and validate a comprehensive model. Then they derived a practical rule of thumb for low leakage cells design.

Usually transistors drain leakage currents are modelled by the combination of two separate phenomena: **diode leakage** and **subthreshold conduction**. In a nMOS the drain to bulk junction current can be written as:

$$I_{db} = -I_s \left[e^{-V_{db}/U_T} - 1 \right] \tag{2.20}$$

where I_s is the diode saturation current and is a process constant proportional to the junction perimeter and area.

The subthershold conduction equation is the Equation 2.18:

$$I_{ds} = I_0 \frac{W}{L} e^{\frac{(\kappa V_{gs})}{U_T}} (1 - e^{-\frac{V_{ds}}{U_T}}) \tag{2.21}$$

thus, the smaller is V_{gs}, the less current flow from the source to the drain.

Since the transistor drain leakage is a combination of these two leakage sources, subthreshold conduction can be made negligible if a small V_{gs} is applied. Resulting in $I_{ds} << I_{db}$.

However, the V_{gs} at which this happens, has two interesting dependencies. First, its a weak function of I_{db}, it means that if I_{db} is reduced, even V_{gs} must decrease accordingly in order to let the subthreshold conduction be negligible.

Second, it depend linearly on the transistor threshold voltage. The smaller is V_{th}, the more negative V_{gs} is required in order to switch the nMOS off. Hence, the use of high threshold voltage devices (n/pMOSHVT), sometimes available in processes, can be beneficial.

Let's start with the first part, the review of existing methods. In order to reduce leakage at the drain terminal, and hence memory degradation if a capacitor is connected to it, two techniques were widely used in past literature. The first consist of a nMOS and a pMOS in parallel forming a transmission gate, see Figure 2.3(a). Since it is made of transistor of opposite type, the leakages of the drain-bulk junctions of each device have opposite directions. If the MOS are carefully sized and matched, these drain-bulk currents can cancel each other and get theoretically zero leakage currents from the drain terminal. This, in practice, is very difficult to get because this nullifying system is very sensible to devices dimensions, mismatches and, not the least, to the MOS operation point.

Another technique consist instead of a single device that has the bulk and the drain terminals tied at the same potential, for instance with a voltage buffer, see Figure 2.3(b). A junction with no difference potential between the two sides doesn't draw any current from the drain terminal. Sure enough, in real world this can happens only if the voltage buffer has zero offset, if not the junction is polarized with the buffer offset voltage and still a small current flow from the drain terminal.

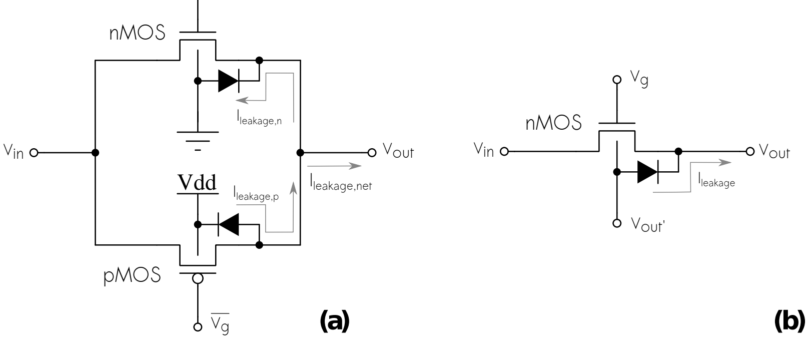

Figure 2.3: Leakage current minimizing techniques. **(a)** nMOS and pMOS in parallel, the two leakage currents counterbalance each other. **(b)** A nMOS cell, here the leakage current is minimized forcing V_{out} to be as close as $V_{out'}$ as possible, via a feedback amplifier.

Technology	Lambda (λ)	Minimum junction Area ($30\lambda^2$)	Scaled reverse-biased junction leakage per unit area	**Estimated** net leakage	**Achieved** net leakage
[μm]	[μm]	[μm^2]	[$\frac{fA}{\mu m^2}$]	[fA]	[fA]
3	1.5	67.5	4 * 0.02	0.22	1.6
2	1	30	3 * 0.02	0.36	1
1.2	0.6	10.8	2 * 0.02	0.017	0.08

Table 2.1: Comparison among estimated and measured leakage reduction performances for a given technology. Table taken from [6].

Despite these technique are well known and widely used in industry, the authors of the paper claims that these past low leakage designs didn't achieve the minimum leakage that they would expect from physical considerations and from math.

In order to prove that, they firstly compared measured leakages found in past papers with their relative processes estimations. They assumed to use low size MOS and that the subthreshold conduction is completely eliminated (reasonable, since it's usually quite easy to achieve). Hence, the only measurable leakage current would be the current junction between drain and bulk. In modern processes reverse-biased leakage per unit area at room temperature is around $0.02 fA/\mu m^2$ [9], this value is decreased by a factor of two to three in the past 15 years as technology improved. Then, they conveniently scaled this parameter according to the process technology and obtainable minimum dimensions. They summarize their analysis in a table, which I partially reproduced in Table 2.1.

That comparison table shows that even the best design circuit has at least 2.5 times larger currents compared to what expected from theory, hence they inferred that optimal design has not been achieved. The authors in [6] claims that:

"These results suggests that additional leakage mechanisms, probably due to the MOS structure, exists and have not been compensated for in these past implementations".

Once addressed the problem, let's analyse the measurements in order to figure out what is unexpected.

In modern processes, biasing the nMOS at $V_{gs} = 0$ means that it's digitally off but if V_{ds} is non zero, I_{ds} will however be larger that I_{db} due to subthreshold conduction.

These considerations are shown in Figure 2.4. In fact, with large V_{gs} the dominant leakage current is the subtheshold conduction (exponential), while with deep negative V_{gs}, the current is leaded by I_{db} and is almost constant. The measured value at which these two currents get equal is around $-200mV$. Hence, the first insight about how to get low leakage in nMOS is to a force a deep negative V_{gs}, resulting $I_{db} >> I_{ds}$.

This can actually be done both by decreasing the gate potential or increase the threshold via the body effect increasing the source potential. They are both feasible and effective for this purpose, but increasing the V_{sb} has the disadvantage that also bias the drain to bulk diode, increasing I_{db}.

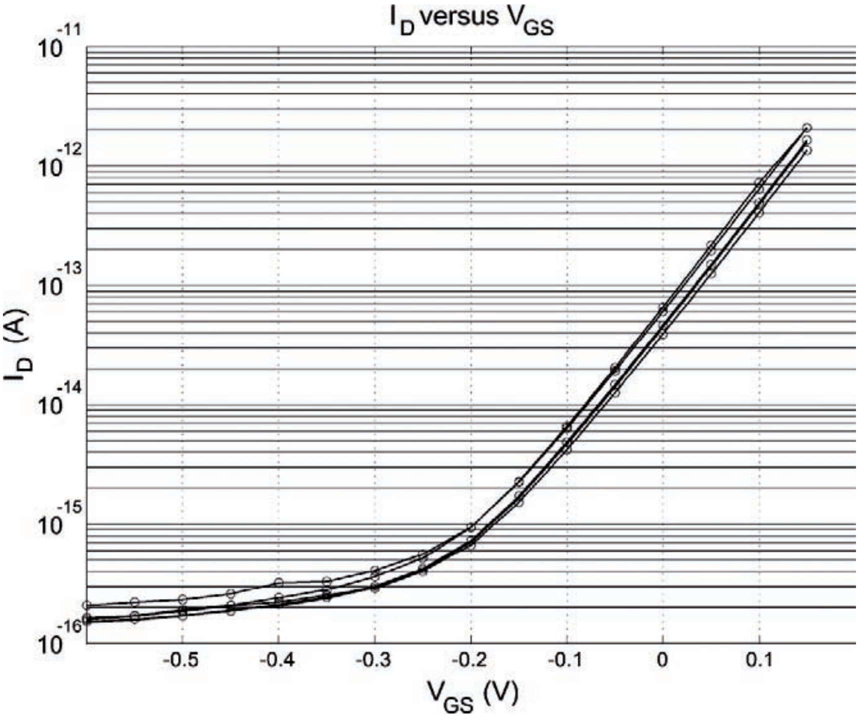

Figure 2.4: Five different minimum-sized nMOS I_d vs V_{gs} curves with $V_s = V_b = 0V$ and $V_{ds} = 150mV$. Plot taken from [6].

Equation 2.21 is valid only in subthershold MOS, i.e. higher than around $-200mV$, according to Figure 2.4. Below that value, that equation is not true any more since the nMOS goes out of the saturation region and enter in the **accumulation mode**. In this region the V_{gs} is so negative that attracts positive holes under the oxide surface and subthershold drain source couple mechanisms are hence negligible. The usual (and wrong) conclusion is than that no source drain coupling is present any more in the device, but the authors carried out intensive measurements on MOS devices in deep accumulation showing that it's not true, and this inaccuracy can be a relevant source of leakage that explains non optimal performance of past circuits.

In Figure 2.5 is shown the unexpected dependencies of I_d to V_{sb} in accumulation mode. This I_d vs V_{gs} plot is obtained with $V_{bulk} = 0$, $V_{db} = 150mV$ and parametrized in $V_{sb} = -50 \div 150mV$. If no accumulation coupling mode were present, the I_d current variation to V_{sb} would be negligible . On contrary, is shown that I_d is actually affected by the source potential. Additionally, as V_{sb} approaches to V_{db}, the conduction current is minimized. This sounds reasonable due to the internal symmetric structure of the MOS.

In Figure 2.6 is shown a I_d vs V_{sb} plot parametrized in $V_{db} = 150 \div 300mV$ given $V_{gb} = -1$ and of course $V_b = 0$. The plot shows that with V_{sb} in the range from $-50 \div +250mV$ the drain current still depend exponentially on the source potential, but is almost constant for $V_{sb} > 100mV$. This residual leakage current is still weakly dependent with the drain terminal due to drain-bulk reverse leakage.

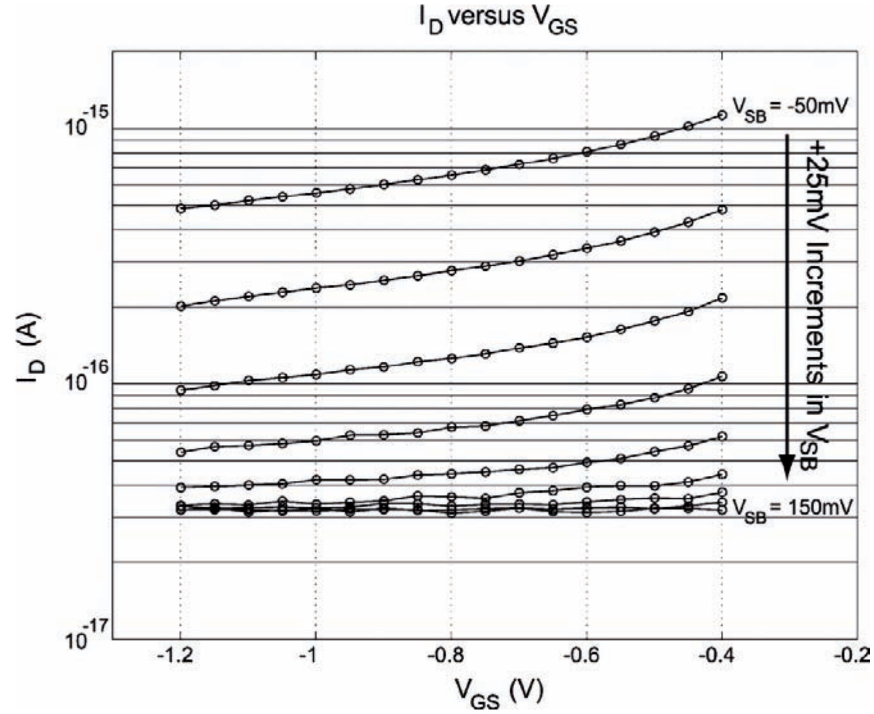

Figure 2.5: Minimum-sized nMOS accumulation-mode I_d vs V_{gs} curves for $-50mV \leq V_{sb} \leq 150mV$, with $V_b = 0V$ and $V_{db} = 150mV$.Plot taken from [6].

These last two plots clearly shows how a source drain coupling is still evident in deep accumulation mode MOS. So, in order to take into account this mechanism, the authors developed an empirical model of the MOS in accumulation mode. Since the coupling effect between drain currents and source potential is exponential, it's natural to model it with an additional parasitic diode D_{sb2} in the MOS structure that only interact with the opposite MOS terminal. While D_{sb1} is the usual source to bulk junction diode.

MOS are symmetric devices, hence the same model can be applied to the drain terminal. Hence, we will have a D_{db2} diode for modelling the accumulation mode coupling with the source and a D_{db1} that only model the drain-bulk junction leakage.

The mechanism that links D_{sb2} and D_{db2} is the diffusion **under** the accumulation layer. This speculation is supported by the fact that in Figure 2.5 the drain current decreases as V_{gs} became more negative. Since the more negative is the gate voltage, the wider is the accumulation mode and hence less source drain coupling is present due to reduction of cross section area of D_{sb2} and D_{db2} junctions.

In this paper were presented only measurements and considerations, no analytical model were developed due to the high complexity of the problem. However, a useful rule of thumb is provided in order to design low leakage switches. Additional experiments were carried later by the same authors [7] and get leakage currents on the order of 5 electrons/second in modern device and differential topology.

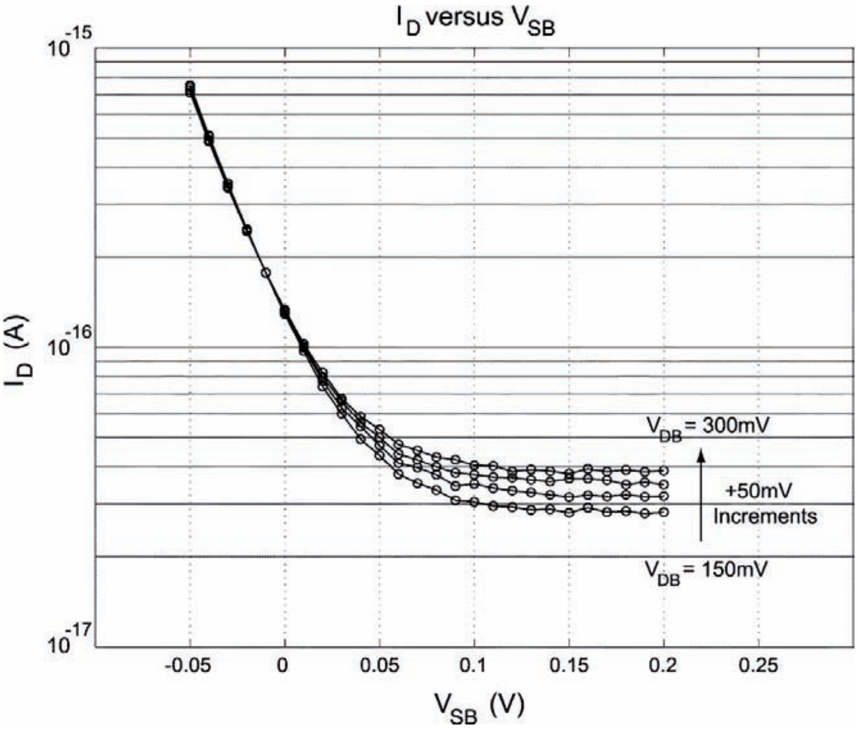

Figure 2.6: Minimum-sized nMOS accumulation-mode I_d vs V_{sb} curves for $150mV \leq V_{db} \leq 300mV$, with $V_b = 0V$ and $V_{gb} = -1$. Plot taken from [6].

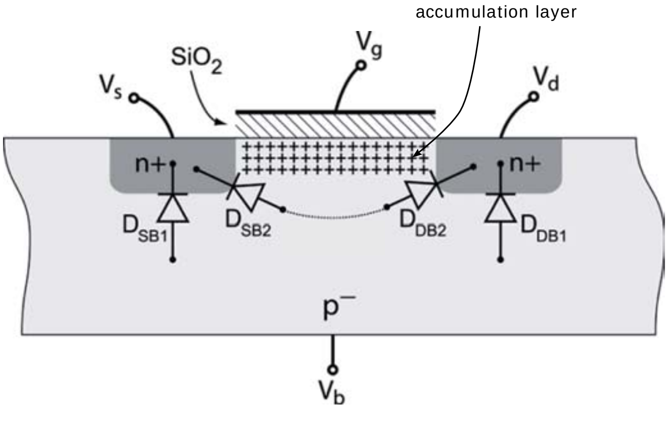

Figure 2.7: The MOS accumulation region model. The drain to source coupling phenomena is modelled by the two additional diodes D_{sb2} and D_{db2}. On contrary D_{sb1} and D_{db1} model the p-n junction leakages. Figure taken from [6].

It's clear form the above discussion that, due to the symmetric nature of the MOS, the source-drain accumulation coupling can be minimized if $V_d = V_s$. In fact, in this situation, the current through D_{sb2} is equal to the current that flow thorough D_{db2} but with opposite direction, hence summing to zero. Finally, since the subthreshold conduction is voided by setting the MOS in accumulation mode, the only relevant drain leakage current would be the one from D_{db1}. This current can be reduced by reverse biasing that junction, yielding the following:

Low Leakage Design rule of thumb (nMOS):

- $V_{gs} < -400mV$, in order to null the subthreshold conduction and biases the MOS in accumulation region.
- $V_{db} = 0$ in order to minimize drain bulk junction leakage.
- $V_{sb} = V_{db}$ in order to minimize drain to source coupling effects.

2.3 The Translinear Principle and log-domain Filters

The *translinear principle* was first introduced by B. Gilbert in the mid '70s, and since that, this technique has been analysed, synthesizes and used by several people. Hence from this, a new paradigm of filters called *log-domain filters* have been developed and formalized by Mulder in [17].

Log-domain filters are time continuous filters that consist of constant current sources, linear capacitors and translinear devices arranged in order to make a translinear loop. The adjective *log-domain* relates to the fact that variables in the circuit are log compressed voltages.

The power and the beauty of this technique is that it is simply based on the exponential laws of certain electronic devices, such as BJT, MOS, IGBT (even combinations of those can be fine) and from a math $log(\cdot)$ propriety. However, to be specific, in this thesis work I will focus only on translinear circuits and log-domain filters made out of MOS devices.

The impact of circuits that rely on that principle is important since it is extremely useful for perform very resources consuming functions such as: multiplications, divisions, squaring, and square rooting in a very efficient way with very few components. In fact, computationally expensive operation such as $exp(\cdot)$ are directly implemented by the device I-V curve itself. Other regular functions, e.g. sum and subtractions, are simply implemented by the KCL on a node. Otherwise, usual techniques that perform such complex functions would require to go into the digital domain and, if not part of a big system with shared resources, the function performances will loose compactness, efficiency and speed.

The modern implementation of *log-domain* filters in CMOS processes is based on the suthreshold MOS in saturation. This is due because this device in this polarization region has exactly the exponential dependence that is required to act as a translinear element and build translinear circuits. In fact, its current-voltage relation was derived in Section 2.1 and described by the Equation 2.19 is rewritten here below for convenience:

$$I_{ds} = I_0 \frac{W}{L} e^{\frac{\kappa V_{gs}}{U_T}} \tag{2.22}$$

The term *translinear* is referred to the fact that the *trans*conductance of the device is *linear* in its collector current. Sure enough this is exactly what happens in the saturated subthreshold MOS:

$$g_m = \frac{\partial I_{ds}}{\partial V_{gs}} = I_0 \frac{W}{L} e^{\frac{\kappa V_{gs}}{U_T}} \cdot \frac{\kappa}{U_T} = \frac{\kappa I_{ds}}{U_T} \tag{2.23}$$

According to Table 1.2, translinear circuits are current-mode circuits, in the sense that their input and output are currents. Inverting Equation 2.22, V_{gs} voltages in a MOS are the $log(\cdot)$ versions of their drain-source currents:

$$V_{gs} = \frac{U_T}{\kappa} ln \left(\frac{I_{ds}}{I_0} \frac{L}{W} \right) \tag{2.24}$$

hence, is clear how translinear circuits compress the voltage dynamic range with a $log(\cdot)$ function. Since voltages can be accommodated in a reduced range, this compression is an additional benefit if the circuit is powered with low power supply, as happens in neuromorphic engineering.

In order to have a translinear circuit, we need a topology in which we can identify a closed loop of n translinear elements. For each of these devices (in our case MOS) they must have the gates and the sources connected at least to one gate or to one source of another translinear element. A topology like that is said to be a **translinear loop** and is shown in Figure 2.8.

Then, in order to analyse the circuit, an arbitrary direction of the loop must be set (let's assume its direction is clockwise). Hence, following the loop direction, if the V_{gs} is positive then the encountered translinear element is said to have a clockwise (CW) direction. On contrary, if the V_{gs} is negative, the translinear MOS is labelled as a counter-clockwise element (CCW).

As long as we travel all around the loop in the direction of the arrow we can apply the KVL and hence write:

$$\sum_{n \in CCW} V_{gs\_n} = \sum_{m \in CW} V_{gs\_m} \tag{2.25a}$$

$$\sum_{n \in CCW} \frac{U_T}{\kappa} ln \left(\frac{I_{ds}}{I_0} \frac{L}{W} \right) = \sum_{m \in CW} \frac{U_T}{\kappa} ln \left(\frac{I_{ds}}{I_0} \frac{L}{W} \right) \tag{2.25b}$$

Additionally, given that all the devices are reasonably at the same temperature, and if the κ of the involved devices are equals, then it reduces to:

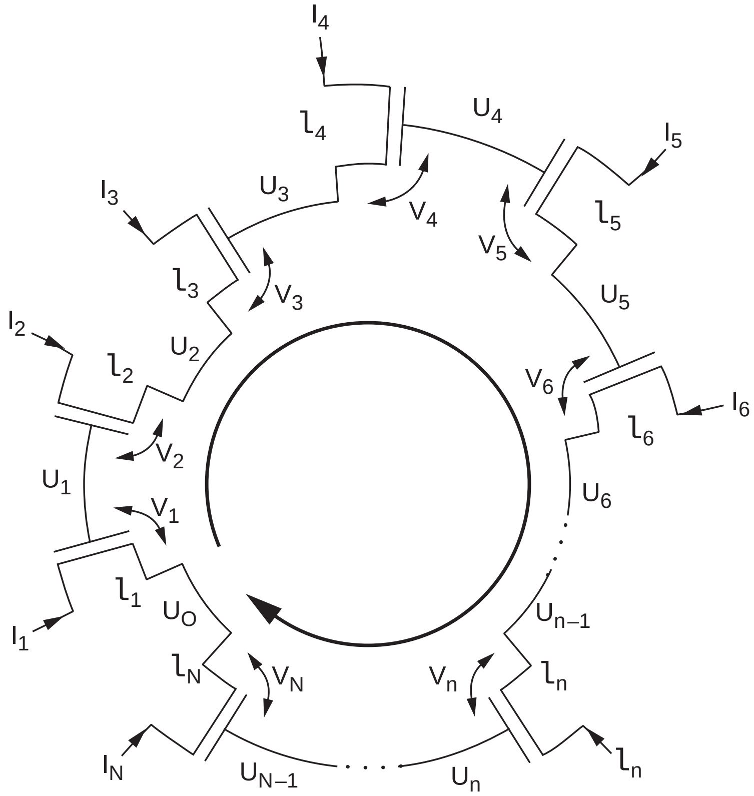

Figure 2.8: A Translinear Loop with MOS elements. Loop in clockwise direction.

$$\sum_{n \in CCW} ln\left(\frac{I_{ds}}{I_0}\frac{L}{W}\right) = \sum_{m \in CW} \ln\left(\frac{I_{ds}}{I_0}\frac{L}{W}\right) \tag{2.26}$$

Recalling that $ln(ab) = ln(a) + ln(b)$ we have that:

$$ln\left[\prod_{n \in CCW} \frac{I_{ds}}{I_0}\frac{L}{W}\right] = ln\left[\prod_{m \in CW} \frac{I_{ds}}{I_0}\frac{L}{W}\right] \tag{2.27a}$$

$$\prod_{n \in CCW} I_{ds}\frac{L}{W} = I_0^{CCW-CW} \prod_{m \in CW} \frac{I_{ds}}{I_0}\frac{L}{W} \tag{2.27b}$$

If the number of CCW and CW elements are equals ($\frac{N}{2}$) then the term I_0 simplifies and the translinear relation finally becomes:

$$\boxed{\prod_{n \in CCW} I_{dsn}\frac{L_n}{W_n} = \prod_{m \in CW} I_{dsm}\frac{L_m}{W_m}} \tag{2.28}$$

This remarkable result is however obtained based on the two non obvious previous assumptions, i.e. the κ of the translinear elements are equals **and** #CCW = #CW. If that holds, than results I_{ds} currents in the translinear elements are related each other to multiplication and divisions and are linearly scaled according to the MOS dimension ratios W/L.

Even though the second assumption is usually quite easy to get, the κ equal assumption is generally tougher. In fact, the MOS body effect results in a κ-dependent exponential constant for the gate and a non-κ-dependent exponential constant for the source in the subtherhold MOS [?]. However, although the body effect is relevant, the *log()* properties still applies for the translinear circuits, but results in undesirable power laws and hence distortion. So, in order to neglect the κ effect, the body effect must be minimized, i.e. the source and the bulk of the MOS must be tied together.

Unfortunately, in a standard CMOS n-well process (like the AMS 0.18μm) only the pMOS bulks are independent from each other since they are made in isolated wells. The nMOS bulks are shared because are the substrate of the die, hence sometimes can be impossible to zeros the bulk effect without forward bias the junctions. However, it is feasible and works well if the sources of the nMOS devices are all at the same potential.

Even the Early effect degrade the tranlinear loop performances. In fact, Equation 2.22 is just an approximation and doesn't take into account this effect. But, if recall the original Equation 2.17, an additional distortion would appear since a weak dependence on V_d is present.

However, both the Early effect and the body effect are usually overshadowed by the mismatch between elements. As already pointed out, mismatch in modern CMOS processes can be huge and it results in important variation from equations if careful design and layout are not performed. Even though the MOS mismatches are countless, since countless are the MOS parameters, the most relevant errors arises from mismatches in threshold values V_{th} and from the mismatch of the transconductances.

A simple yet instructive example of log-domain filter made out nMOS devices is the one shown in Figure 2.9. It is basically the well known current mirror with the addition of a capacitor. If the two nMOS are biased in subthreshold saturation they acts as translinear elements and hence a translinear loop can be identified. For the circuit analysis let's set the loop direction to be CW (as highlighted in the figure) and given that the first encountered terminal of M_1 is its source, M_1 is then a labelled as CW element. On contrary, since M_2 is first encountered at the gate terminal, M_2 is hence a CCW element.

Is shown that #CCW = #CW, so the first assumption is verified. Since this is a very simple circuit, even the κ assumption is

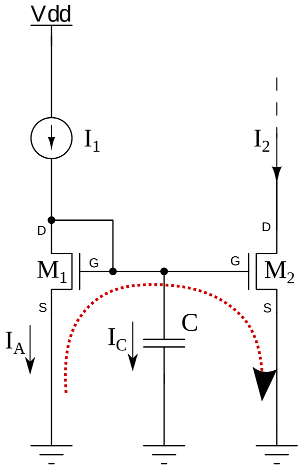

Figure 2.9: A simple *log-domain* filter made out of current mirror and a linear capacitor.

verified. In fact, the bulk terminals of the two device can be connected at their source terminals cancelling the body effect and thus getting equals κ. Recalling the capacitor equation we can write:

$$i_c(t) = C\frac{dV_c}{dt} = C\frac{dV_{gs2}}{dt} = \frac{CU_T}{\kappa i_2(t)}\frac{di_2(t)}{dt} \tag{2.29}$$

and from KCL holds $i_1(t) = i_A(t) + i_C(t)$, then applying the translinear principle and combing these two Equations yields:

$$i_A(t) = i_2(t) \tag{2.30a}$$

$$i_1(t) - i_C(t) = i_2(t) \tag{2.30b}$$

$$i_1(t) = i_2(t) + \frac{CU_T}{\kappa i_2(t)}\frac{di_2(t)}{dt} \tag{2.30c}$$

$$\tag{2.30d}$$

then Laplace transforming and setting $\tau = \frac{CU_T}{\kappa I_2}$ it results:

$$I_1(s) = I_2(s)(1 + \tau s) \longrightarrow \frac{I_2(s)}{I_1(s)} = \frac{1}{(1 + \tau s)} \tag{2.31a}$$

Giving a LPF first order log-domain filter.

This simple yet meaningful example gives an insight about translinear loops an log-domain filtering. Second order effects, such as threshold voltage deviations, mismatches, early effects are easily understandable by this simple example. However, all of that degradates the circuit performances in terms of static errors (current ratio) or in dynamic errors (distortions).

In this discussion about translinear circuits we assumed currents that carry informations are always positive. This is not always true in real circuits, hence arises the problem how to represent such a negative signals. This issue is usually overcome by implementing differential translinear loops, as a result the effective signal is the difference between two positive currents, and thus can be negative. Otherwise you can keep the single ended circuit and add an offset current as happened in the mirror example.

In this section I gave an overview and an example about how the translinear principle is really useful for make filters. In fact, it can be done by simply adding capacitors in one translinear loop node. A further and more interesting example about low pass filter is the Differential Pair Integrator circuit, shown in section 2.4.

2.4 The Differential Pair Integrator Circuit

The Differential Pair Integrator is a *log-domain* circuit developed by C. Bartolozzi at INI in 2005. This circuit goal is to emulate synapses in silicon that exhibits compactness, low power consumption but still a wide control of the dynamics. The DPI is hence one of the basic building blocks of the neural architecture implemented in the MN256r1 chip.

The basic DPI circuit is shown in Figure 2.10, it consist of four nMOS and two pMOS and one capacitor. As will be clear later, M_{thr} sets, via its gate terminal, the DPI threshold while M_τ and C_{syn} affect the circuit time constant, i.e. the dynamics. M_{in} is the input nMOS and acts as a current source according to the input signal (in our case a spike). The maximum input generated current is weighted by the V_w bias of M_w. M_{syn} is the output pMOS that acts as a current generator, which current magnitude is proportional to the input current I_{in} and to the circuit parameters.

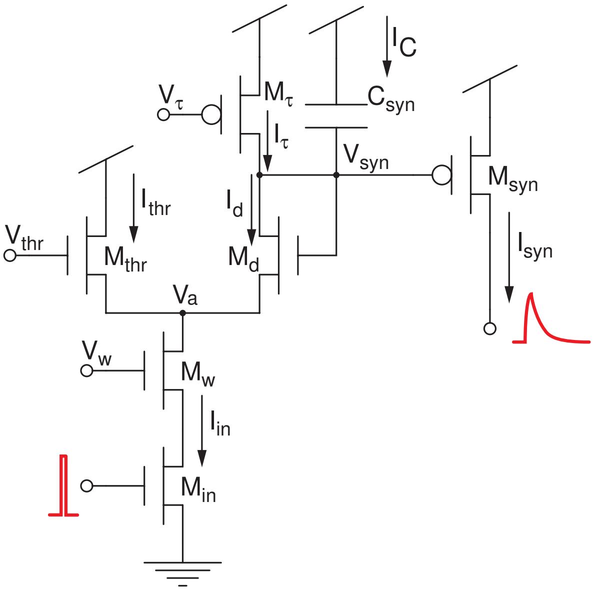

Figure 2.10: The DPI schematic.

If we are at steady state condition, the current I_{in} is constant, and it's split in two components that goes through M_{thr} and through M_d. Since in steady state condition the capacitor doesn't let current flows, so $I_C = 0$ and I_τ is equal to I_d of M_d. Hence I_{thr} is equal to $I_{in} - I_d$. In order to satisfy this last equation V_{gs} of M_{thr} is then set by the device physics and is not a variable. Since V_{thr} is externally set, its source potential V_a is at:

$$V_a = V_{thr} - \frac{U_T}{\kappa} ln \left(\frac{I_{thr}}{I_0} \frac{L}{W} \right) \tag{2.32}$$

Note that, ideally this source imposition doesn't interfere with the current I_{in}, since the early effect is negligible.

Now we are in the situation where M_d has both the current imposed (I_τ by M_τ) and the source terminal potential V_a (due to M_{thr} previous consideration). So its gate terminal will be set by the circuit accordingly to the usual formula 2.19. The M_d diode connected nMOS allows it because its drain potential is ideally independent from its current. It results that, at steady state a certain potential V_{syn} affected by V_{thr} and by I_{in}, sets the output current I_{syn}.

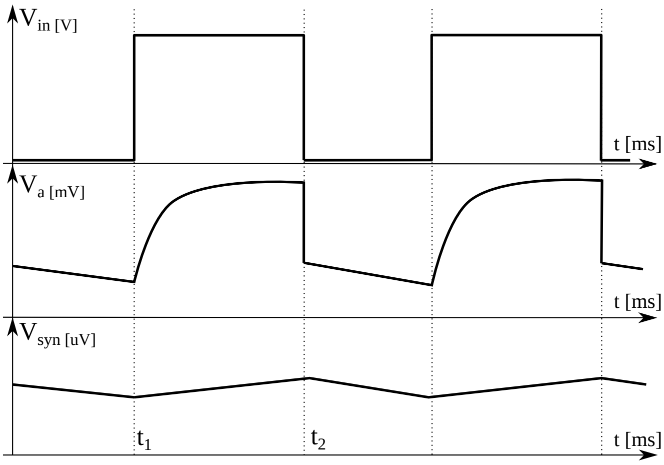

Figure 2.11: Qualitative time plot of meaningful voltages in the DPI schematic draw upon simulation plots. Note the different y-axis scales.

But, what happens if a variation in the input current occurs? Let's now assume we are at the steady state and suddenly increase the input current. Such additional current must flow somewhere according to the KCL, i.e. in M_{thr} or in M_d or both. However, these nMOS gates are fixed, both from external generator or by the capacitor inertia. Hence, the only way to allow the increase in current is to decrease their source potential V_a down to a local equilibrium is reached. This quickly happens at t_2 (see Figure 2.11) and it results in an increase in both the currents, namely I_{thr} and I_d.

Since I_d is statically given by the fixed I_t, the additional required current must be drained out from the capacitor. However, if the capacitor drain currents, it results in a flow of electrons and then in a decrease of potential of the capacitor. Hence, V_{syn} potential is not fixed any more and goes lower. If V_{syn} is lowering, I_τ is not affected, but I_d will and hence a smaller currents flow though M_d. In order to satisfy the KCL $I_{in} = I_{thr} + I_d$ the common node V_a gets slightly lower (see after t_2). Since the decrease of V_a and of V_{syn} are set by different dynamics, their difference V_{gs} of M_d will converge and end up to an equilibrium point is reached.

Instead, if we suddenly decrease the input current, happens that V_a increase in order to reduce I_{thr} and I_d and satisfy the KCL at that node. But now $I_d < I_\tau$. Hence, the current difference at t_1 flows into the capacitor giving $I_C = I_d - I_\tau$. As before, while the current goes into the capacitor, it raise up V_{syn} and hence increase I_d reaching the equilibrium point where $I_d = I_\tau$. However, this charging phase is much slower than the discharging phase due to a bigger current that flows in the capacitor. As contrary as before, while V_{syn} increases, the output current I_{syn} gets lower.

This was just an intuitive explanations in order to to really understand the circuit, an analytic derivation based on the translinear loop is given here below.

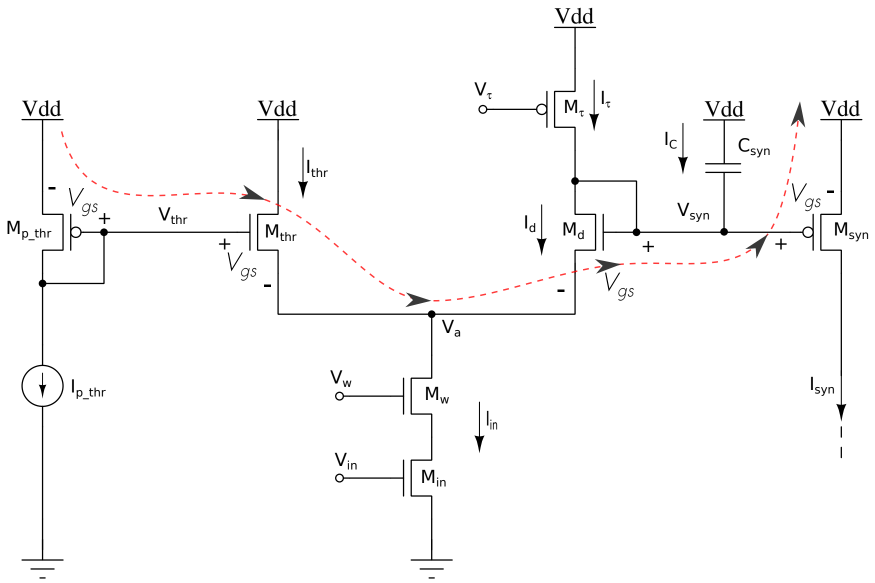

Figure 2.12: The DPI schematic with the translinear loop in evidence. Note the additional pMOS on the left (M_{p_thr}) added for convenience.

Relating to Figure 2.12 let's start with the net and the elements equations:

$$I_{syn} = I_0 e^{\frac{\kappa V_{syn}}{U_T}} \qquad I_{in} = I_{thr} + I_d \qquad I_d = I_\tau + I_c \tag{2.33a}$$

$$I_c = C\frac{dV_{syn}}{dt} = C\frac{U_T}{\kappa I_{syn}}\frac{dI_{syn}}{dt} \tag{2.33b}$$

If the circuit elements are biased in subthreshold saturation region, hence the translinear loop applies (see dotted line in Figure 2.12). The resulting relation is then:

$$I_{p\_thr} \cdot I_{thr} = I_d \cdot I_{syn} \tag{2.34}$$

Combining Equations 2.34 and 2.33a we can write:

$$I_{p\_thr}(I_{in} - I_\tau - C\frac{U_T}{\kappa I_{syn}}\frac{dI_{syn}}{dt}) = I_{syn}(I_\tau + C\frac{U_T}{\kappa I_{syn}}\frac{dI_{syn}}{dt}) \tag{2.35a}$$

and dividing both terms by I_τ and defining $\tau = \frac{CU_T}{\kappa I_\tau}$

$$\tau\left(1 + \frac{I_{p\_thr}}{I_{syn}}\right)\frac{dI_{syn}}{dt} + I_{syn} = \frac{I_{p\_thr} I_{in}}{I_\tau} - I_{p\_thr} \tag{2.36}$$

This is a first order non linear differential equation that can't be solved analytically. However, if $I_{in} >> I_\tau$ the term I_{p_thr} on the right side of Equation 2.36 can be neglected. And if $I_{syn} >> I_{p\_thr}$ even the term $\frac{I_{p\_thr}}{I_{syn}}$ can be dropped. This last assumption true since even though $I_{syn} = 0$ at the beginning, it will increase monotonically and meet the $I_{syn} >> I_{p\_thr}$ condition over time.

Hence, we can rewrite Equation 2.36 as:

$$\boxed{\tau\frac{dI_{syn}}{dt} + I_{syn} = \frac{I_{in}I_{p\_thr}}{I_\tau} \longrightarrow \frac{I_{syn}(s)}{I_{in}(s)} = \frac{I_{p\_thr}}{I_\tau}\frac{1}{1+s\tau}} \tag{2.37}$$

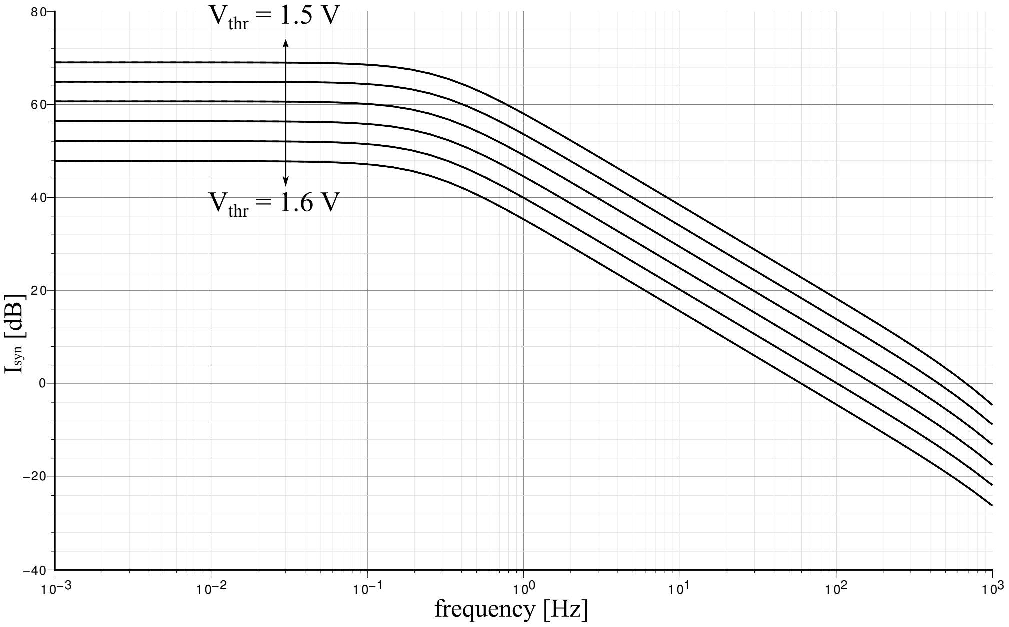

Figure 2.13: The DPI AC response, parametrized with V_{thr}, it changes the circuit gain but not the time constant. All transitors are sized: $W/L = \frac{1\mu m}{1\mu m}$.

As described in Section 1.1 meaningful signals in neuromorphic engineer are spikes. Even though the action potential spike dynamic is shown to be quite complex, is proven [2] that a first order approximation is sufficient for compute and process information in large neural networks. Once again, the computational power is results form highly interconnected neuron rather than a faithfully reproduced dynamics.

A spike can be obtained by applying a digital impulse to the DPI input. If that impulse is narrow enough, i.e. one or two orders of magnitude below the DPI time constant, happens that the rising time of the I_{syn} signal is really fast since the exp in a neighbour of the origin can be approximated as a steep line. But, since the spike is narrow and ends soon, the DPI doesn't reach the steady state point and goes back to its resting value. Hence, given a narrow digital impulse, the output is an analogue signal that resemble the neuron action potential. See Figure 2.15 and Figure 2.16 for the temporal dynamics of the DPI.

As already pointed out, this circuits implements only a basic action potential. In MN256r1 other circuits are responsible for inter-spike dependent dynamics. One of those circuits that

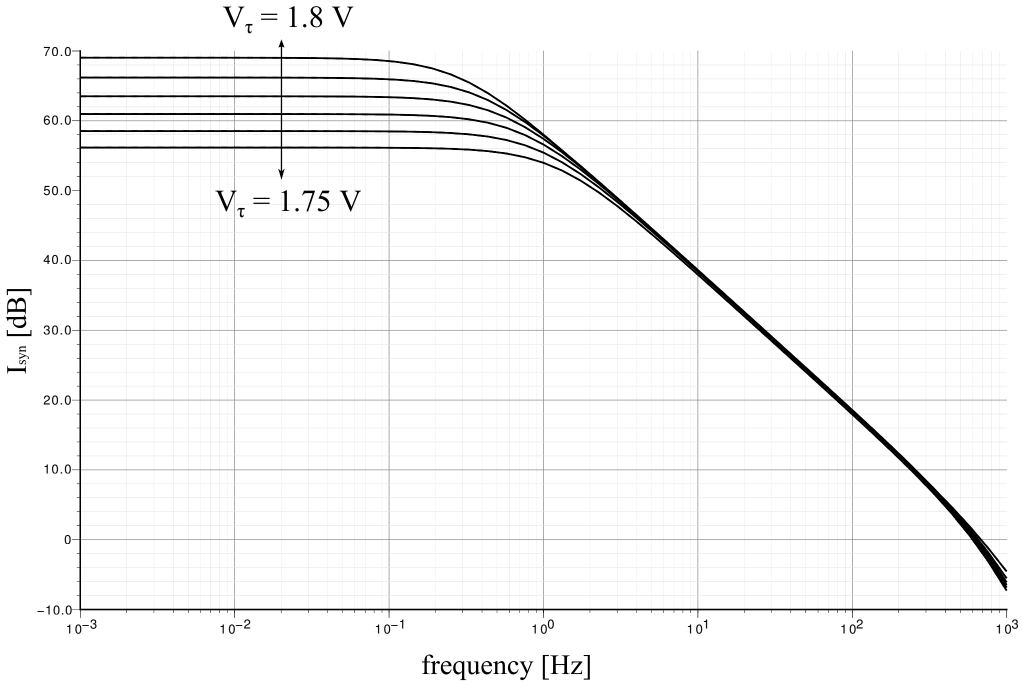

Figure 2.14: The DPI AC response, parametrized with V_τ, it changes both the cut-off frequency and the circuit gain.

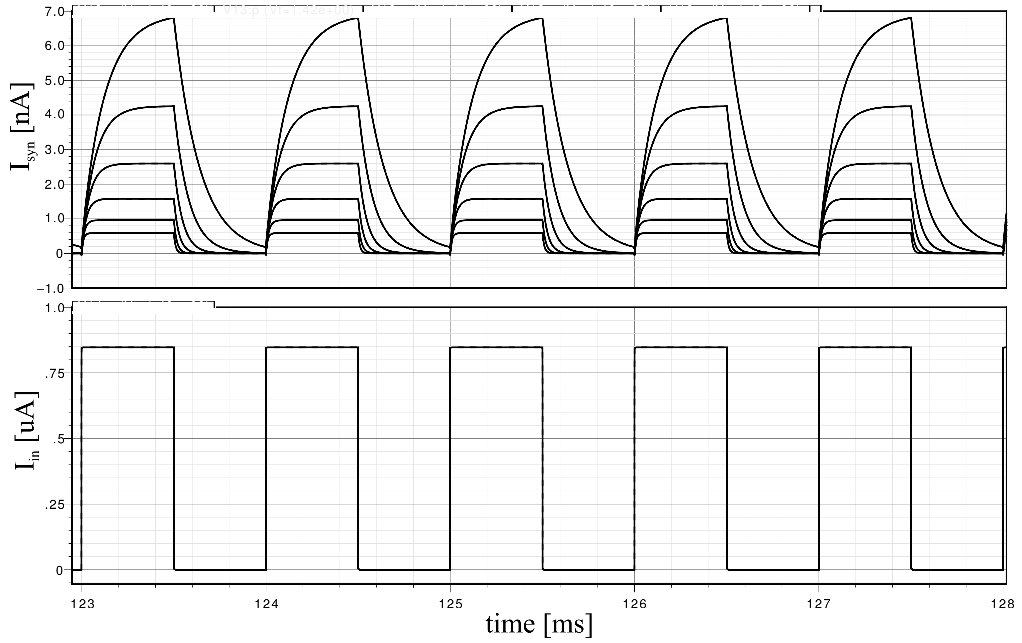

Figure 2.15: The DPI response of a square wave duty 50% input signal, parametrized in V_τ. All transistors sized: $W/L = \frac{1\mu m}{1\mu m}$

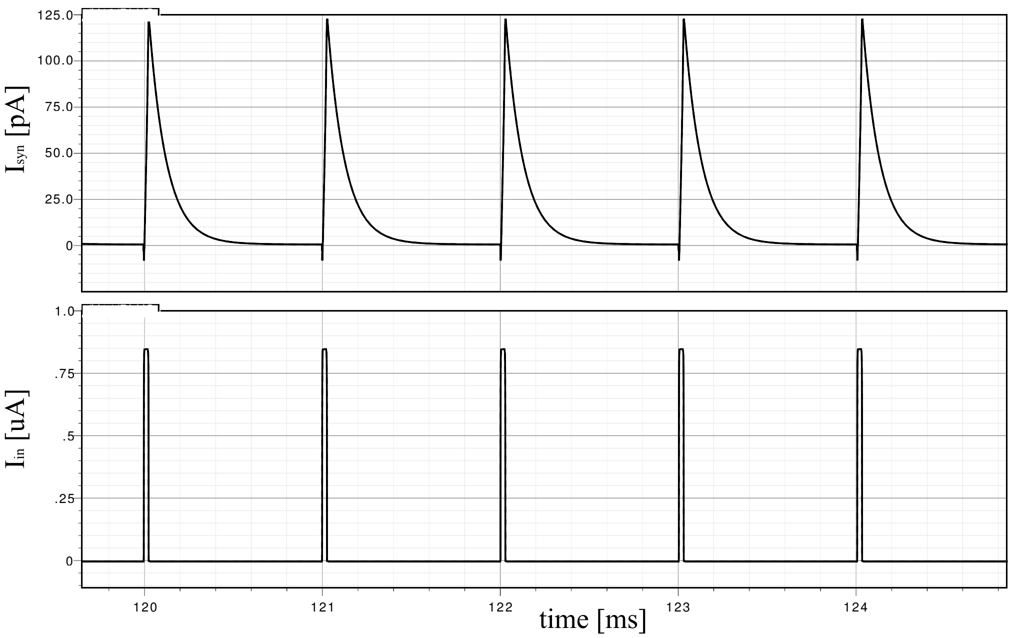

Figure 2.16: The DPI response of a pulse input signal. The output shape recall the biological spike of Figure 1.3.

will be interfaced directly to the DPI is my homoeostatic circuit that implement the homoeostatic plasticity.

2.5 The Winner-Take-All Circuit

The Winner Take All circuit is a continuous time analogue signal processor invented by Lazzaro et al. in the late 80's. It consist of two MOS cells and exhibits a very compact, parallel and highly modular architecture for comparing signal magnitudes. Mathematically speaking, it actually implement the $max(I_{in1}, I_{in2}, ...I_{inN})$ function among N inputs, and select the "winning" output letting it flow the whole bias current I_{tail}. Hence the name Winner Take All (WTA), in the sense that the most rugged signal wins over the others.

A basic two cells WTA schematic is shown in Figure 2.17. There we have a cell made by M_2 and M_4 arranged as a local feedback and connect one input I_{in1} to the circuit. M_3 and M_5 are the second input cell and process the I_{in2} signal. Any further inputs can be added by simply insert additional two nMOS cell for each of the required inputs.

M_1 acts as current generator I_{tail} which current represent the maximum output current. I_{tail} can be seen as the "prize" for the "winner" among the competing input signals. Additionally, it sets the circuit gain and directly affect the power consumption.

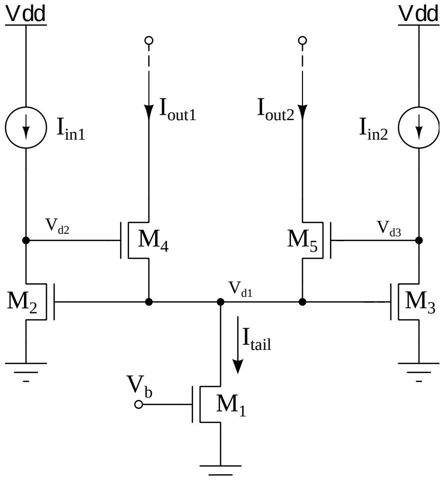

Figure 2.17: A two input Winner Take All schematic.

Except where we intentionally want asymmetric and weighted behaviour for the inputs, the WTA cells usually consist of nMOS of equal sizes. The only nMOS that can be differ from the other is of course M_1. Indeed, given a symmetric circuit, if the two inputs I_{in1} and I_{in2} are equals then the tail current set by M_1 is hence equally split on I_{out1} and I_{out2}.

But what happens if the two inputs are remarkably ($I_{in1} >> I_{out2}$) different?

If this happens while M_2 an M_3 are in saturation region, both try to satisfy Equation 2.19, and a conflict arises. In fact, these transistors would like to set their gate terminals at different values, but since they belong to a shared node V_{d1}, it can't happen. The value of V_{d1} is set by the larger of the gate voltages between M_2 and M_3, i.e M_2 (the one with the highest input current).

Hence, given that M_3 has the V_{gs} set by M_2, in order to be $I_{in2} = I_{ds2} << I_{in1}$ must happens that M_3 changes its bias point and goes out of saturation region and falls into the triode region. In fact, recalling Figure 2.2, with fixed V_{gs}, we see that I_{ds} considerably lowers if V_{ds} decreases and let M_3 goes into the triode region. Since V_{d3} lowers and V_{d1} is fixed by the other input branch, the current through M_5, i.e. I_{out2} severely decreases. Given the KLC at V_{d1} node, is true that $I_{tail} = I_{out1} + I_{out2}$, and then we can conclude that $I_{out1} \approx I_{tail}$.

The "winning" input has hence been "rewarded" let him taking the whole pot, i.e. I_{tail}.

Note that an increase of I_{out1} will increase V_{d2}, but since the I_{ds} to V_{ds} slope is negligible in saturation compared to the one of triode region, this variation won't heavily affect V_{d1} and won't at all influences large signals operations.

But, what happens if the difference between the inputs is small? Recall Figure 2.2, we see that in saturation, the slope of the MOS is not zero due to nonidealities and second order effects (Early voltage). That behaviour can be included in the nMOS subthreshold model by the following formula:

$$I_{ds} = I_0 \frac{W}{L} e^{\frac{\kappa V_{gs}}{U_T}} \left(1 + \frac{V_{ds}}{V_E}\right) \tag{2.38}$$

where V_E is the Early voltage.

Starting from the equilibrium point ($I_{in1} = I_{in2}$) we slightly increase the value of I_{in1} by δ_I. Hence, its drain voltage will increase by:

$$\delta_V = \frac{\delta_I}{I_0 \frac{W}{L} e^{\frac{\kappa V_{gs}}{U_T}}} V_E \tag{2.39}$$

Since V_{d2} is also the gate M_4, I_{out1} will be amplified by an amount proportional to e^{δ_V}. Hence, due to KCL at node V_{d1}, the output current I_{out2} must decrease by the same amount. The gain of the competing mechanisms between two input signals is $\frac{\delta_V}{\delta_I}$ and is directly proportional to the Early voltage V_E and inversely proportional to the tail current I_{tail}.

A simulation of a two input WTA fully symmetric circuit is plotted in Figure 2.18 with two different values of I_{tail}. The bigger is the bias current, the steeper is the plot in the x-axis zero neighbourhood. As explained before, with big differential inputs, the output currents saturates at the bias current value.

A variation of the schematic shown in Figure 2.17 is depicted in Figure 2.19. The difference of this last circuit is that one output current is doubled via a proper sized mirror and forced to flow in a branch shared with a constant current source which value is I_{tail}. If the input signal I_{in1} is different compared to the I_{in2}, hence $2I_{out1}$ is either close to zero or to $2I_{tail}$. Given such a different of imposed currents, the results is that the drain terminals of the output transistors M_8 and M_9 will goes to an appropriate condition in order to satisfy KCL in the branch. Since the gain of the WTA is high, hence the V_{out} swing ranges from almost ground up to the power supply. However, a particular equilibrium condition arises when $I_{in1} = I_{in2}$ and yields a $V_{out} = V_{out}^*$, which value lays in between gnd and Vdd.

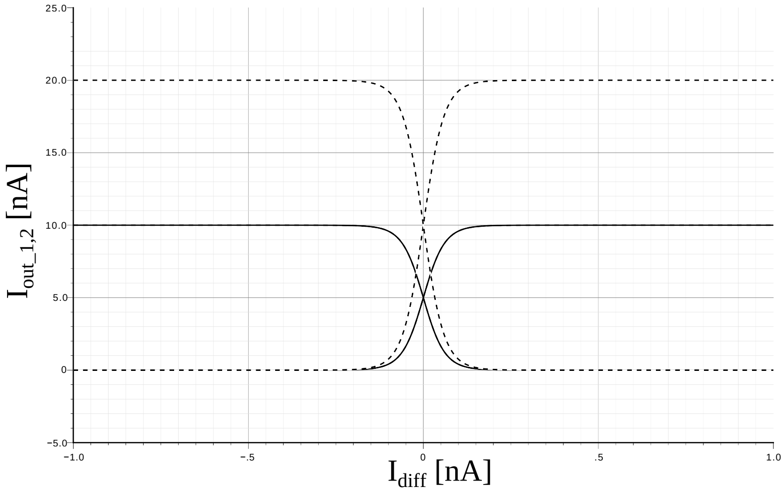

Figure 2.18: The WTA static plot. All nMOS are sized $\frac{W}{L} = \frac{1\mu m}{500nm}$, $I_{tail} = 10 \div 20nA$, $I_{comm\_input} = 10nA$ and $I_{differnetial} = \pm 1nA$.

The equilibrium point V^*_{out} depends on the MOS sizes and on the common input current value.

This structure has a behaviour that resemble a digital current-voltage logic gate with an additional equilibrium point somewhere in the middle of the voltage swing, or it can seen as an analogue comparator. However, it can be modelled as follows:

$$V_{out} = \begin{cases} 0 & \text{if } I_{in1} < I_{in2} \\ V^*_{out} & \text{if } I_{in1} = I_{in2} \\ Vdd & \text{if } I_{in1} > I_{in2} \end{cases}$$

The temporal dynamics of this modified WTA circuit is simulated and plotted in Figure 2.20.

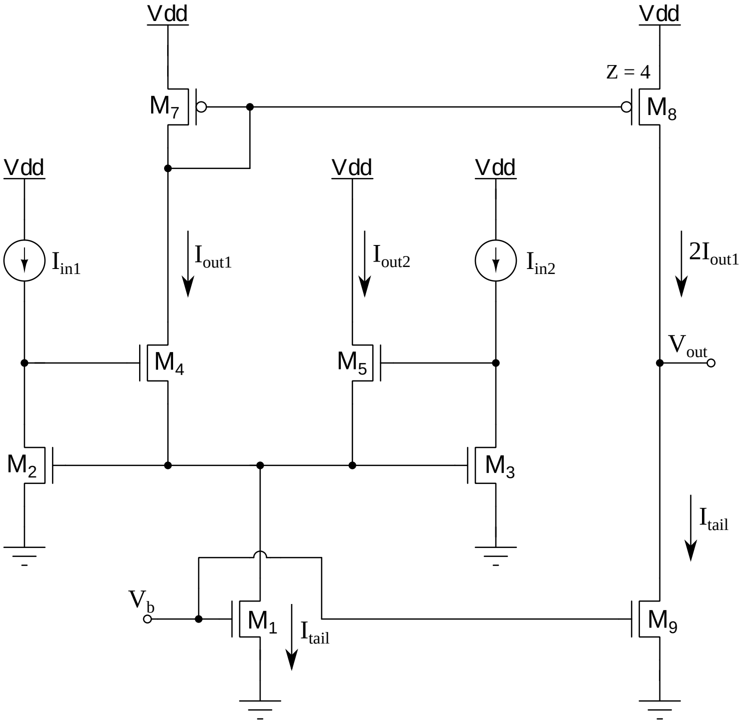

Figure 2.19: A modified WTA with voltage output. All the MOS are sized $Z = 2$ except for M_8 that is sized $Z = 4$.

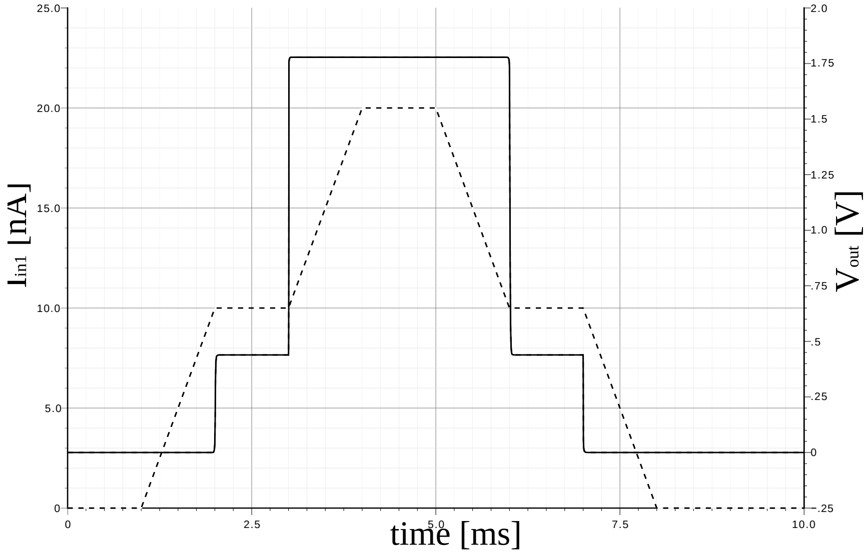

Figure 2.20: The enhanced WTA time simulation with $I_{tail} = 10nA = I_{ref}$. The solid line is the voltage output while the dashed one curve is the current input. All MOS have $Z = 2$ but M_8 that has $Z = 4$.

CHAPTER

3

THE FILTER DESIGN

In this chapter I firstly restate the homeostatic principle in a engineering way providing a transition from the biological domain to the silicon domain. The key point to that is to adopt a standard AGC architecture. Then I moved on topology design considerations of LPF needed in the AGC with an implementation level emphasis. As already announced, I tried to report in this thesis work even some intermediate solutions that didn't works fine for our application but still emphasize circuits limitations. In particular, I report here two intermediate design (Design 1 and Design 2) that were not suitable to obtain very small cut-off frequencies. The final design (Design 3) that gives nice results is presented at the end of the chapter and will be the adopted solution that successfully implements really long time constants on silicon.

3.1 The Automatic Gain Control

The Automatic Gain Control (AGC) is a circuit that adjust the gain of an amplifier in order to have a constant output amplitude in face of amplifier input variations. This type of circuit is wide known in industry and is commonly used in hard disk drive read channels, medical and multimedia systems CCD sensors, wireless receiver and so on. It is basically a negative feedback loop and one way to implement this automatic gain correction is to sense the output signal of the amplifier and modify the gain of the amplifier accordingly. This is the **feedback AGC** implementation, the opposite to that is the feedforward implementation that conversely senses the input of the amplifier instead of the output. However, the key element of the AGC circuit (both feedback and forward) is a LPF that averages the amplifier output signal (or input in the forward realization) and provides smooth and low frequency signals suitable to be used as gain control commands. This value is used to set the gain of the amplifier in a way that the amplifier gain increase if the output signal strength is too low, and the amplifier gain decrease if the sensed output strength is too high.

From this brief discussion and recalling Section 1.4 is clear that **the AGC is the engineering implementation of the homeostatic principle present in real neurons.** In fact, the neuron can be seen as an amplifier with the pre-synaptic current I_{in} as the input and with the spike frequency rate as output f_{out}. This relation can be assumed to be linear and the slope of the gain is set by internal parameters of the circuits and by V_{thr} (I_{p_thr}) of the DPI.

Thus, to implement the homeosatic principle in silicon, the proper solution is to sense the output firing rate of the neuron, compare it with a frequency reference, process the difference with a LPF and feed back this processed signal in the synapses changing its V_{thr}. This solution is depicted by a block diagram in Figure 3.1.

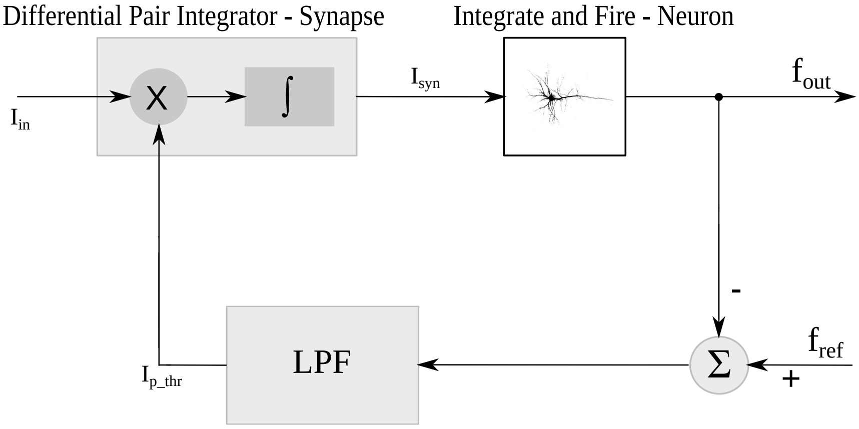
Differential Pair Integrator - Synapse
Integrate and Fire - Neuron
I_in
X
∫
I_syn
f_out
-
f_ref
+
Σ
LPF
I_p_thr

Figure 3.1: Block diagram of the AGC loop applied to an artificial neuron of MN256r1 chip. Each neuronal instance consist of the cascade of a DPI synapses and an integrate and fire neuron.

However, would be computationally expensive to extract the frequency from an analog signal consist of spikes (such as f_{out}), and compare it to a fixed frequency f_{ref}. An implementation to that would be to add an additional filter that converts f_{out} fast spikes into a low-frequency current or voltage which value is proportional to the neuronal firing ratio. This additional filter, plus to the AGC-LPF and the DPI integrator, results in a more complex implementation third order loop, which stability is not guaranteed.

These problems arises since the information is taken at the very output of the neuron. But, can we do something to avoid it?

Recalling Figure 1.2, since $f_{out} = f(I_{syn})$ is assumed to be linear in small variation around the stable point, f_{out} is a constant gain scaled version of I_{syn}, hence $f_{out} = hI_{syn}$, where

$h \in \mathbb{R}^+$. So, f_{out} and I_{syn} carry the same information. Given that observation, the AGC that implements the homeostatic principle is not required to sense f_{out} since the information can be instead recovered from $u(t)$. This is a great advantage because simplifies the AGC circuitry implementation. The schematic block of the simplified AGC that implements the homoestatic principle in silicon is given in Figure 3.2.

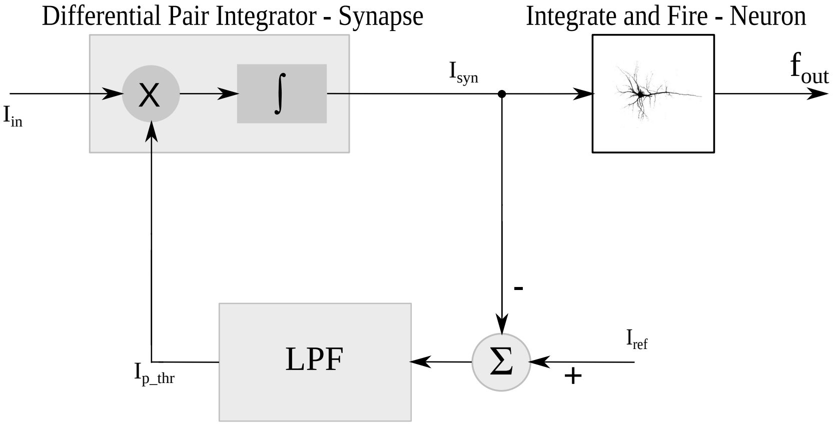
Differential Pair Integrator - Synapse
Integrate and Fire - Neuron
I_{in}
X
∫
I_{syn}
f_{out}
-
I_{ref}
+
Σ
LPF
I_{p_thr}

Figure 3.2: Block diagram of the simplified AGC implementation of the homoestatic principle on MN256r1 neuronal instance.

In this figure, I_{ref} is the reference current at which the system state, when the AGC dynamics are over, will converge. In fact, if $I_{ref} - I_{syn}$ is not zero, this difference is processed by the LPF and its output value affects the DPI gain via I_{p_thr} (i.e. V_{thr}). This will change I_{syn} in a direction (negative feedback) that results in $I_{ref} = I_{syn}$, thus giving I_{p_thr} constant. To note that now we can't directly set the firing rate of the neuron, but we can only set its synaptic current. This means that, in order to get the desired f_{out}, I_{ref} must be set accordingly, i.e. $I_{ref} = \frac{f_{outref}}{h}$. In practice, this is not an issue since I_{ref} is a manual bias.

If the assumption $f(I_{syn}) = h$ is not true, a difference between the desired output average firing rate and the real one is affected by and error, that can be only static if $f(t) = h^* \neq h$ with $h^* \in \mathbb{R}$, or dynamic if $f(I_{syn})$ can't be approximated with a linear equation. However, high precision in f_{ref} is not required in our application, both concerning its absolute value and its variance between neuronal instances in the chip.

As already highlighted before, the homeostatic principle dynamics must be kept several order of magnitude slower that the learning mechanisms dynamics, i.e. the DPI dynamics. This point is a key specification and almost all the effort in the design path were concentrated here. However, this practical consideration is very important even while concerning about the AGC stability. In fact, referring to Figure 3.2, the AGC loop consist of a second order system. One pole is associated to the DPI, and the other comes form the LPF. These two poles would give the system in a potential instability, since the phase margin can be very close to 0°. Fortunately, this is avoided in our circuit since the big difference in the two cut-off frequencies yields poles very distant to each other, let the system to be modelled as a dominant pole system. In fact, given a fixed gain system, only the dominant pole sets the bandwidth of the loop and the phase margin is around 90°. Since the influence of the second pole is way negligible if they are spaced by frequency decades, it doesn't degrades the phase margin.

3.2 The Core Idea

In order to obtain ultra long time constants using small area (i.e. small capacitors), a current generator of tiny (femto to atto Amperes) currents is required to charge and discharge the filter capacitor. Usually, a current generator consist of a MOS device with the gate and the source potentials set by Equation 2.19 in order to get the desired current I_d. However, this implementation has important limitations concerning the minimum obtainable current I_d due to unwanted leakages mechanisms that easily dominates at such low currents regimes.

As already anticipated in Section 2.2, tiny currents can be obtained by applying the **rules of thumb** for a low leakage cell design. In fact, measurements on [6] and [7] show that such a technique can gives a leakage, hence a current, down to 5/electrons per second in modern CMOS processes. In their papers, the authors show an application of their results in a sample and hold cell and in a switched capacitor filter only. Indeed, they never attempted to use their results to build a tiny current generator.

From the structure they presented, no straightforward derivations of a femto to atto Amperes current generator with one floating terminal can be obtained. In fact, is not easy to generate such low leakage currents and, at the same time, let those tiny currents flow into a capacitor without being affected by other device currents. This behaviour is mandatory in order to set temporal dynamics and hence build filters.

A possible implementation of a tiny current generator that I developed on purpose for this thesis, based on the *low leakage design rules of thumb*, is shown in Figure 3.3.

That circuit comprises a pMOS polarized as low leakage current source via two ideal amplifiers A_1 and A_2 that actively satisfies two bullets (the 2nd and the 3rd) of the *low leakage design rule of thumb*. In fact, A_2 forces the V_{db} voltage to be zero, reducing the drain bulk I_{db} leakage current. Meanwhile, A_1 forces the drain and the source terminals to be at the same potential, cancelling the source drain accumulation mode coupling.

The gate terminal is tied at V_{dd} (or at the maximum voltage available in the chip) hence, if $V_c = V_d = V_b = V_s$ ranges from ground to $V_g - 400mV = 1.8 - 0.4 = 1.4$ then, the pMOS is biased in accumulation mode and the subthreshold conduction is voided. In this scheme, the only current that can flow out from the drain terminal of the pMOS goes into one capacitor terminal. Since this terminal is not tied to any fixed potential, charging or discharging of the

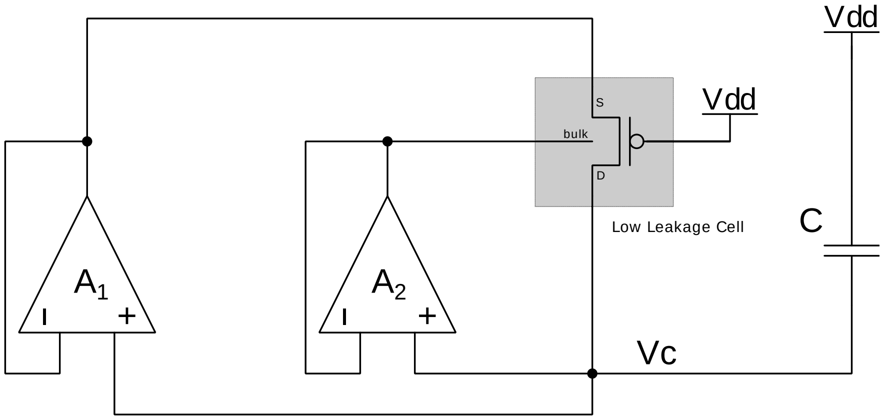

Figure 3.3: A Low current generator with one floating node structure. This is inspired by concepts and measurements of [6] and [7].

capacitor can happens according to the sign of I_d pMOS current.

If the pMOS is biased into accumulation mode and the others Low Leakage Cell (LLC) rule of thumb are satisfied, we see from Figures 2.5 and 2.6 that the tiny I_d current is V_{gs} invariant and almost constant if small variations of V_{ds} are experienced by the device. The fact that I_d is V_{gs} invariant is great since the source, the bulk and the drain terminals are imposed by the capacitor stored charge and will surely vary according to the filter dynamics in normal circuit operations.

However, the LLC pMOS first bullet requirement ($V_{gs} > 400mV$) is not guaranteed by any topological condition and can incidentally be violated if not taken into account by the designer at a higher level. So, if happens that $V_{gs} < 400m$, then the I_{ds} term exponentially increases due to subthreshold conduction that considerably increase I_d current. Sure enough, this is a situation that we must avoid by any means at regimes operations.

Even though the low leakages rule of thumb holds both for nMOS and pMOS, in my circuit, I choose to implement the Low Leakage Cell with a pMOS for two reasons:

The first is that I needed to have independent access and control to the bulk terminal of the MOS in order to reduce the leakages. As already pointed out in Section 2.3, our process AMS $0.18\mu m$ is a n-well process, hence only pMOS can be made in a separate well. On contrary, all nMOS shares bulks, prevent to the possibility to control it independently. Additionally, is beneficial that the pMOS has lower conductivity compared to nMOS devices. This is due because holes mobility is less than electrons mobility $\mu_h < \mu_e$.

Additionally, I want to stress out that the V_c node, that is the one that carries information, is really sensible and must be handled with care. It means that any electrical devices connected to it must exhibit really high impedance seen from that node. This is mandatory and is an important condition in order to not insert additional leakage sources that would otherwise vanish all the LLC pMOS improvements and degrades circuit performances in term of cut-off frequency.

Sure enough, this particular care has been applied in the case of my circuit of Figure 3.3. In fact, the node V_C, is connected only to the two gates of the amplifiers. This connection can statically leaks out out currents in the order of few atto ampere (simulated), due to second order phenomena such as tunnel effect into and through the gate oxide and from the injection of hot carriers from substrate to the gate oxide [8]. However, the V_c node can be dynamically affected by capacitive coupling. This is a serious problem if in the circuit there are digital like (control) signals or fast analogue dynamics, such as spikes. Actually, except for one control $V_{control}$ signal that will be introduced and explained later, that is not our case due to careful design.

Given that, one should avoid to connect anything else on that sensible node in order not to interfere and introduce additional leakage mechanisms. But, if needed, in order to close the AGC, the only thing allowed is to connect to V_c a gate of a MOS not topologically close to signals with fast dynamics.

However, even though dangerous signals are distant from the topology point of view, this may not be guaranteed in the layout. So, additional care must be performed during this last design phase with shields and guard rings.

Note that the drain and the source terminals of the low leakage cell are somehow arbitrary in the scheme of Figure 3.3. In fact, according to the definition, they are relative to each other and depend according to their voltage magnitude. With ideal amplifiers, they are impossible to label since $V_d = V_s$. However, in real world we will deal with real amplifiers made out by OPAMPs and hence with all their limitations, such as finite gain, distortion, small output resistance, additional poles, and so on. In this case, important nonidealities affect the voltage difference between the noniverting terminals of the OPAMP, V_{in+}, and its output, V_{out}. This difference $V_{diff} = V_{in+} - V_{out} = V_{finite\_gain} + V_{offest}$ is the resulting combination of the finite gain of the OPAMP V_{finite_gain} and of the offset voltage V_{offset}. That offset voltage is mainly due to internal device processes mismatches, Early effects and other nonidealities.

The usual low frequency model of a real OPAMP is

$$V_{out} = (V_{in+} - V_{in-})A \tag{3.1}$$

where the gain (A)is finite. If the OPAMP is feedbacked in unity gain voltage follower can write:

$$W = \frac{V_{out}}{V_{in+}} = \frac{A}{1 + A} \tag{3.2}$$

Note that if $A = \infty$ hence we get the ideal OPAMP relation $V_{out} = V_{in+}$. But if it is finite $A < \infty$ hence is true that $V_{out} < V_{in+}$.

However, even though this relation, we are not sure weather the V_{out} is higher or lower compared to V_{in+} due to the additional term V_{offset}. That therm is a poorly controllable value (within certain boundaries) and can be both positive or negative for each instance of the OPAMP. Hence, the tiny accumulation mode current I_{ds} in the pMOS can flow both upwards or downwards in the pMOS according to the sign of V_{diff} of OPAMP A_1. Additionally, V_{diff} of OPAMP A_2 set the magnitude and the sign of I_{db} leakage current. Hence, the sign and the magnitude of the final current I_d, responsible to alter the capacitor charge, is affected by the nonidealities of both the two OPAMP A_1 and A_2.

Previously, I stated that $V_s \simeq V_d \simeq V_b$ must range from ground to $V_{dd} - 0.4mV = 1.4V$ in order to let the pMOS be in accumulation mode. Hence, this LLC pMOS consideration gives only an upper bound limit if OPAMPS are ideal. However, other important limitations of the OPAMPs are its input and output voltage swing. In fact, if a on purpose advanced design has not been performed, such as [24], OPAMP suffer of applicable input voltage limitations. In fact, recalling that the input stage of a simple OPAMP is usually made by a single differential pair. Let's consider the nMOS differential pair case, hence a minimum DC V_{in+} is required in order to let it be biased in the correct region (saturation). It results in a lower bound OPAMP working condition. On contrary, if the differential pair consist of pMOS then the working condition gives an upper bound.

As last point I also want to note note that two distinct OPAMPs in this particular circuit are useless. In fact, they shares both the input terminals $V_{in+_{A1}} = V_{in+_{A2}} = V_C$ and the output terminals $V_{out_{A1}} = V_{out_{A2}}$, hence only one OPAMP would be sufficient to the task. However, the circuit is presented in this form since in later sections this A_1 and A_2 merging can't be performed any more and two separate and independent OPAMPs are thus required.

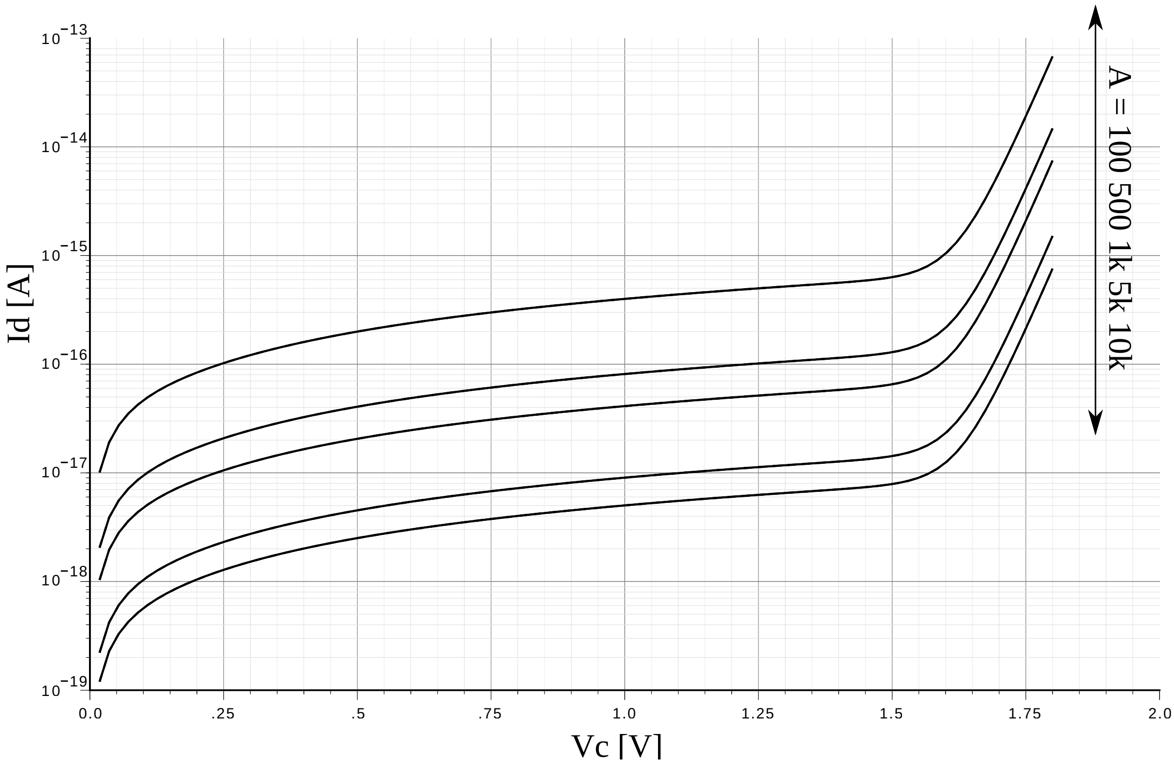

Figure 3.4: I_d vs Vc plot of Figure 3.3, parametrized in A that is the gain of both OPAMP A_1 and A_2. Note that, if V_c is high (above 1.5V) the drain current increases exponentially. This means that the subthrasheold conduction is becoming dominant, as shown in real chip measurements of Figure 2.4. Additionally, the current magnitude is modulated by the OPAMP gain. In fact, the lower is the gain, the more nonideal is the OPAMP and the value $V_{in+} - V_{out}$ increase, increasing I_d as well.

3.3 Filter Design 1 - The DPI

Since in the AGC loop a LPF is required, one straightforward solution is to combine the DPI circuit of Section 2.4 with the Low Leakage Cell of Section 2.2.

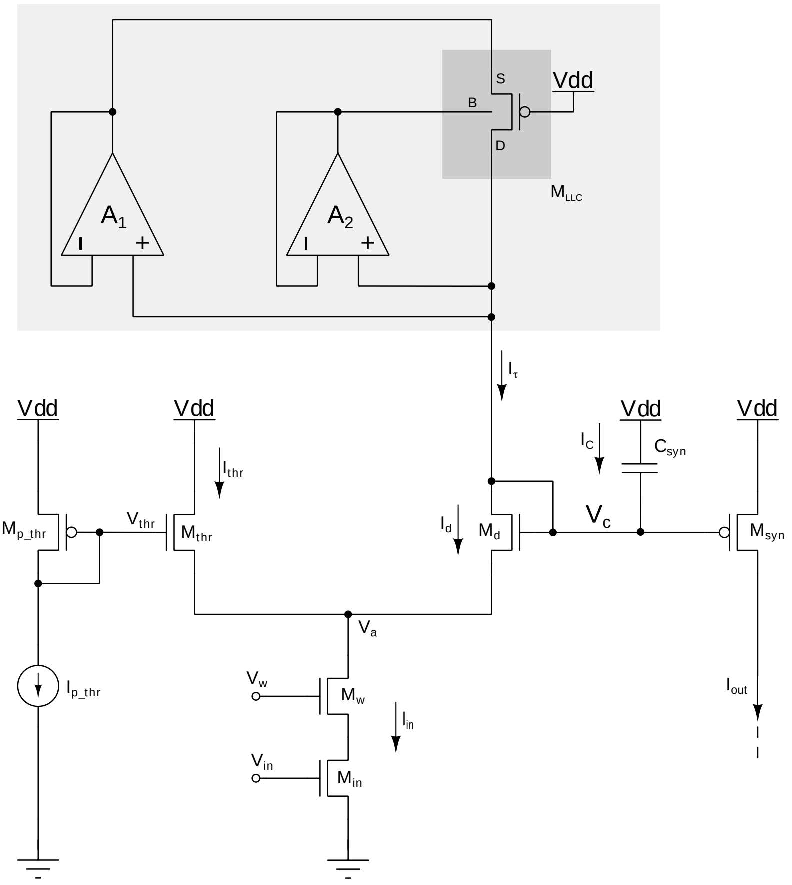

Figure 3.5: The schematic of the first LPF design attempt. The I_τ current generator is here implemented by the LLC.

In fact, the idea is to replace I_τ with the LLC, connecting the V_C terminal with the gate and drain of M_d of Figure 2.10. According to the DPI dynamics equation $\tau = \frac{CU_T}{\kappa I_\tau}$, this topology should give very long time constants. Unfortunately this is not true due to second order effects and nonidealities of the DPI circuit that were not taken into account in Section 2.4 analysis. In fact, the bulk terminal of transistor M_d is connected to ground and is shared among all nMOS devices in a n-well process. Hence, a positive V_{db} is set and drains out a leakage current I_{db} in the order of $50fA$. This contribute is negligible if I_τ is well above it, i.e. if $I_\tau >> I_{db\_M_d}$ and the time constant is then set only by I_τ. But, if I_τ is comparable with $I_{db\_M_d}$ hence the time constant is altered by the drain bulk junction current and effectively limits the DPI performances. In the sense that the capacitor is not solely charged by I_τ but is dominated by $I_{db\_M_d}$ and $\tau = \frac{CU_T}{\kappa I_\tau}$ doesn't hold any more.

Note that, even though in a two well process would be allowed to get a separate p-well for the nMOS M_d, I_{db} can't be minimized by forcing $V_d = V_b$. In fact, that would yields to forward bias the bulk to source junction of M_d.

A time plot simulation of the Filter Design 1 is shown in Figure 3.6. Is clear that the circuit time constant increase as I_τ lowers, but at a certain point it stops and stay constant. This happens between curves that corresponds to $I_\tau = 10fA$ and $I_\tau = 5fA$. In fact, below $I_\tau = 10fA$, the charging current is not dominated by I_τ any more but on contrary by $I_{db\_M_d}$, that is insensitive the to the value of I_τ.

Additionally, as I_τ lowers, even the DPI gain lowers, according to the DPI transfer function of Equation 2.37.

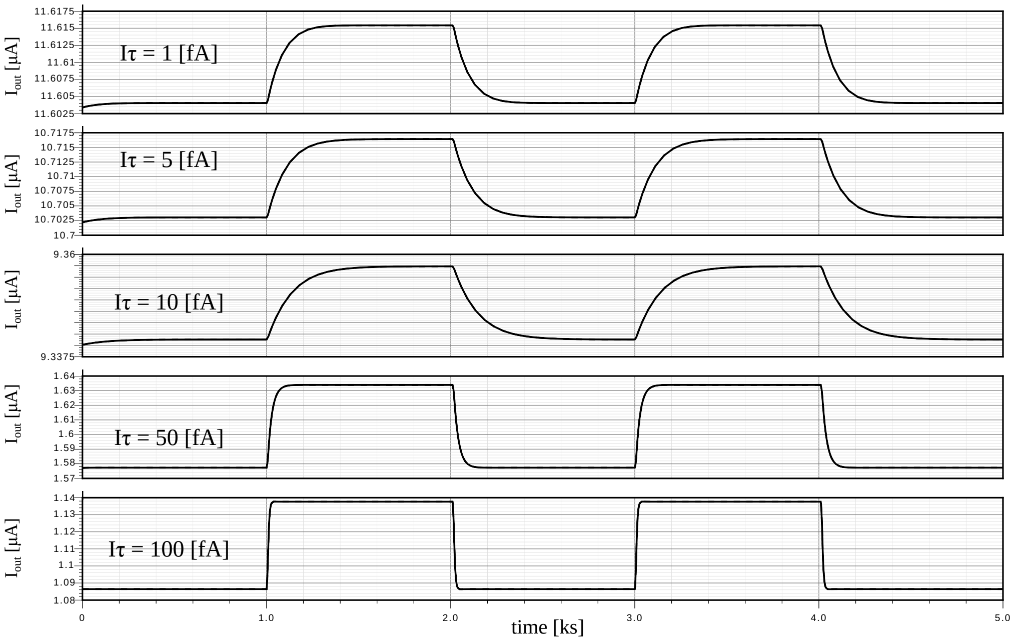

Figure 3.6: Time simulation of the DPI I_{out} given a I_{in} square wave from $5nA$ to $6nA$. $I_{p\_thr} = 1nA$, $C = 10pF$. The simulated $I_{db\_M_d} \approx 50fA$.

3.4 Filter Design 2 - The Classical log-domain

The problem of the previous design came up because drain and source terminals of a critic MOS were connected at different potentials. Still remaining in the log-domain filtering, a different topology based on [9] has been analysed. This architecture is shown in Figure 3.7 and comprises four pMOS, three current generator and the usual capacitor.

If all transistors M_1 to M_4 are biased in subthreshold saturation region, we can analysed the circuit behaviour by applying the translinear loop technique. This topology has the M_3 pMOS, that is functionally equivalent of M_d for the DPI. However, in this topology, M_3 has the $V_{sb} = 0$ since the pMOS bulk is accessible and can hence be connected to the drain. This connection can be done for each of the four pMOS that comprises the translinear loop (M_1 to M_4), giving no Body effect and resulting in κ parameter equal between pMOS. As already emphasized before, if the κ term is not constant in all translinear elements, distortion increase.

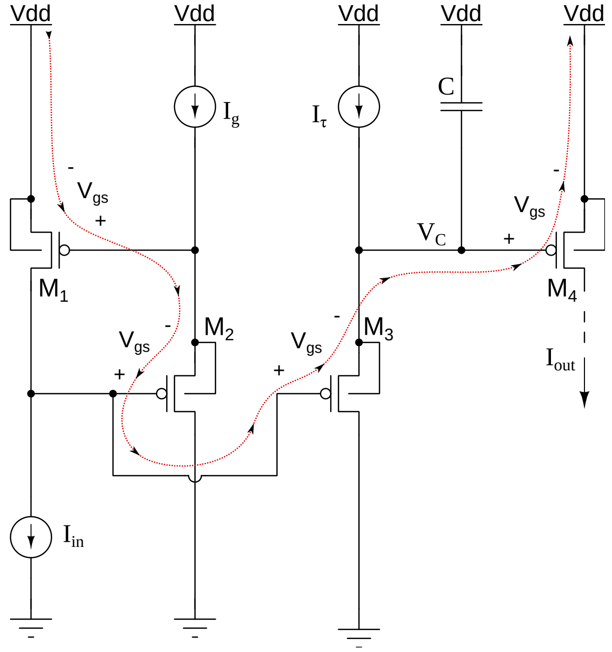

Figure 3.7: The schematic of the second design attempt.

Let's now derive its transfer function. Following the red translinear line and applying KVL is true that:

$$V_{gs_1} + V_{gs_2} = V_{gs3} + V_{gs4} \tag{3.3}$$

Hence, for simplicity assume that $\frac{W}{L} = Z$ is equal between the pMOS. Applying Equation 2.28 gives:

$$I_{in} I_g = (I_\tau + i_C) I_{out} \tag{3.4}$$

recalling that $i_C = \frac{CU_T}{\kappa} \frac{1}{I_{out}} \frac{dI_{out}}{dt}$, Equation 3.4 becomes:

$$I_{in} I_g = I_\tau I_{out} + \frac{CU_T}{\kappa} \frac{dI_{out}}{dt} \tag{3.5}$$

by labelling $\tau = \frac{CU_T}{\kappa I_\tau}$ and Laplace transform it yields:

$$\frac{I_{out}(s)}{I_{in}(s)} = \frac{I_g}{I_\tau} \frac{1}{(1 + s\tau)} \tag{3.6}$$

Note the Equation 3.6 is equivalent to Equation 2.37 of the DPI with $I_g = I_{p\_thr}$.

Since the source and bulk terminals of M_3 are tied together (Figure 3.7), ideally no I_{sb} current would flow in that junction. However, that is not enough since I_{db} is reverse biased and, due to KCL, I_{db} current must flow out from the bulk and hence through the capacitor, altering its charge value. An idea in order to fix this problem is presented in Figure 3.8 an consist of an unity gain OPAMP that senses the source voltage of M_3 and hence sets the bulk of M_3 accordingly. If the OPMAP is ideal the condition $V_s = V_b$ is once again satisfied. However, the difference from before, is that the drain to bulk current that flow in the bulk now comes from the output stage of the OPAMP instead of from the capacitor.

So, even this circuit was a nice candidate in order to implement the low pass filter with the desired ultra long time constants. A schematic of the simulated circuit is shown in Figure 3.8 and once again consist of the union of schematic of Figure 3.7 and Figure 3.3.

Unfortunately, even though that problem is now fixed, another issue arise. In fact, we see that at steady state condition the tiny current I_τ flow in transistor M_3. Since I_τ is the current that flow both on M_{LLC} and on M_3, they must have similar operation point. I.e. for very small I_τ happens that M_3 is in accumulation mode too and hence the filter doesn't behave as predicted by Equation 3.6 (the assumption that all the MOS are in saturation region doesn't hold any more).

Simulations of Figure 3.9 shows that performances of Design 2 are better in term of time constants compared to Design 1, but still exhibits problem and can't work if I_τ is too low.

This is reasonable since the translinear loop is not exploitable any more. So, I concluded that also this second design can't work properly as a filter due to intrinsic problems that can't fixed with this topology.

From these attempts I see that the challenging part of the project was not only to generate really low currents, as I thought at the very beginning of my thesis work, but even being able to exploit it in an effective low cut-off filter.

These, and several others circuits, were object of my studies and simulations. However, none of the conventional filters that I took into account were suitable for ultra low currents and hence ultra long time constants. These intrinsic hindrances urged me to develop a new solution from scratch. The proposed Design 3, that is the one by which I got the best results, is discuss here below in Section 3.5.

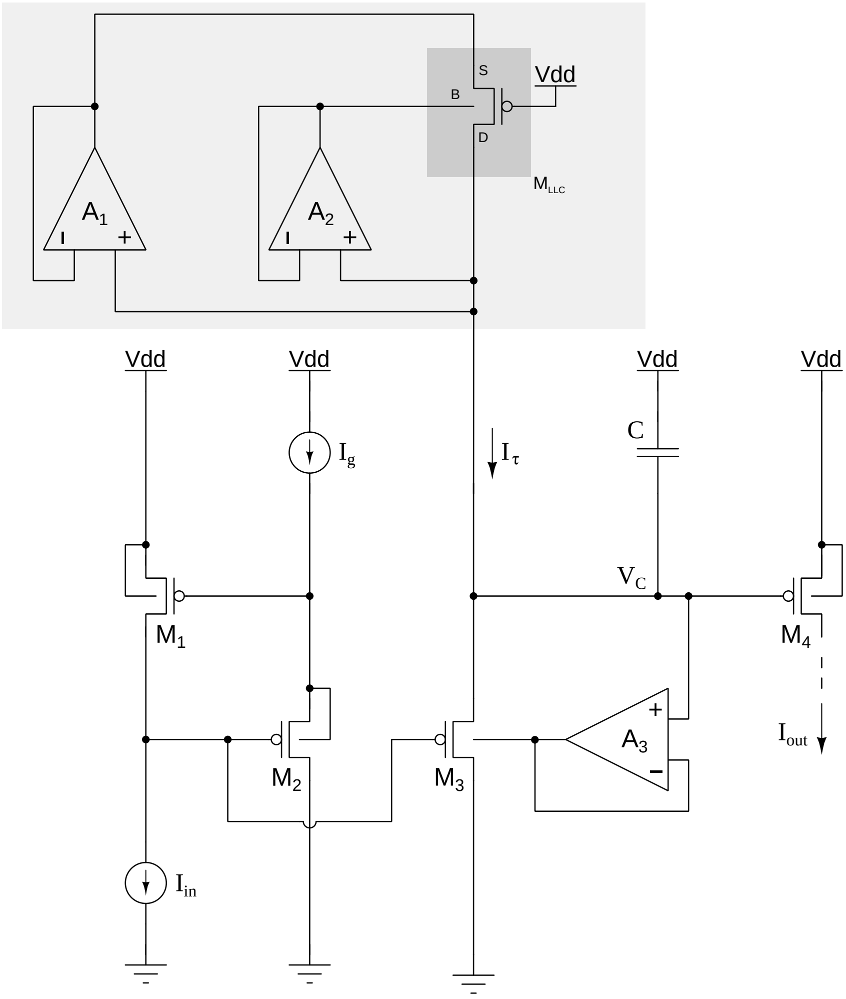

Figure 3.8: The schematic of the second design attempt. The I_τ current generator is here implemented by the LLC.

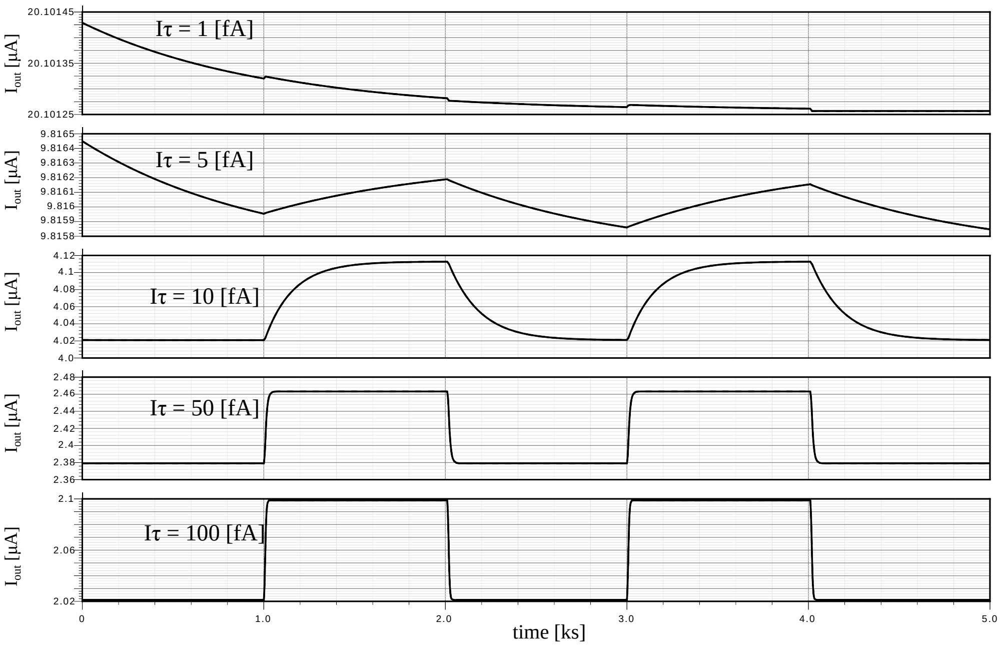

Figure 3.9: Time simulation of the output current of the circuit of Figure 3.8. I_{out} is obtained by applying a I_{in} square wave from $5nA$ to $6nA$. $I_{p\_thr} = 1nA$, $C = 10pF$.

3.5 Filter Design 3 - The Unbalanced Structure

Since previous filter designs yielded poor performances that didn't meet the specifications, a new approach to the problem was developed. I decided to obtain ultra long time constants based on a different architecture that, as far as I know, has never been used before.

In fact, even though the circuit of Figure 3.3 of Section 3.2 worked fine, the issue arised while I tried to connect it to other circuits in order to obtain a filter. Given this issue, my idea was to develop a filter-like circuit built around the circuit of Figure 3.3 instead of try to insert it into an existing filter topology.

Recalling Figure 3.3 and simulation plot in Figure 3.4, if the OPAMPs are completely ideal, the circuit is onset. But, if the OPAMP A_1 has offset (no matter its nature), then a small current can flow upwards or downwards in the LLC pMOS according to the magnitude and sign of the OPAMP offset. Hence, my idea was to somehow get control of that offset and used it for set the sign of the current in the LLC pMOS and hence in the capacitor.

As a results, due to tiny currents, this unbalanced structure would results in a circuit with ultra long time constants where the sign of $\frac{dV_C}{dt}$ is decided according to a control signal ($V_{control}$) that on purpose put offset the OPAMP.

Given this consideration, a circuit that implement this idea is shown in Figure 3.10. The LLC part is exactly the same as presented in Figure 3.3, but this structure has additional external circuitry that set small voltage differences V_{ds} across the LLC according to the $V_{control}$ value.

In fact, the nMOS are sized as follow: $M_1 = \frac{1\mu m}{1\mu m}$, $M_2 = \frac{2\mu m}{1\mu m}$ and $M_3 = \frac{3\mu m}{1\mu m}$ and, via the input terminal $V_{control}$, is possible to decide **the sign and the amount of the pMOS I_d current.** This external circuitry has the same results as modify the OPAMP internal offset of Equation $V_{diff} = V_{in+} - V_{out} = V_{finite\_gain} + V_{offest}$ of Section 3.2.

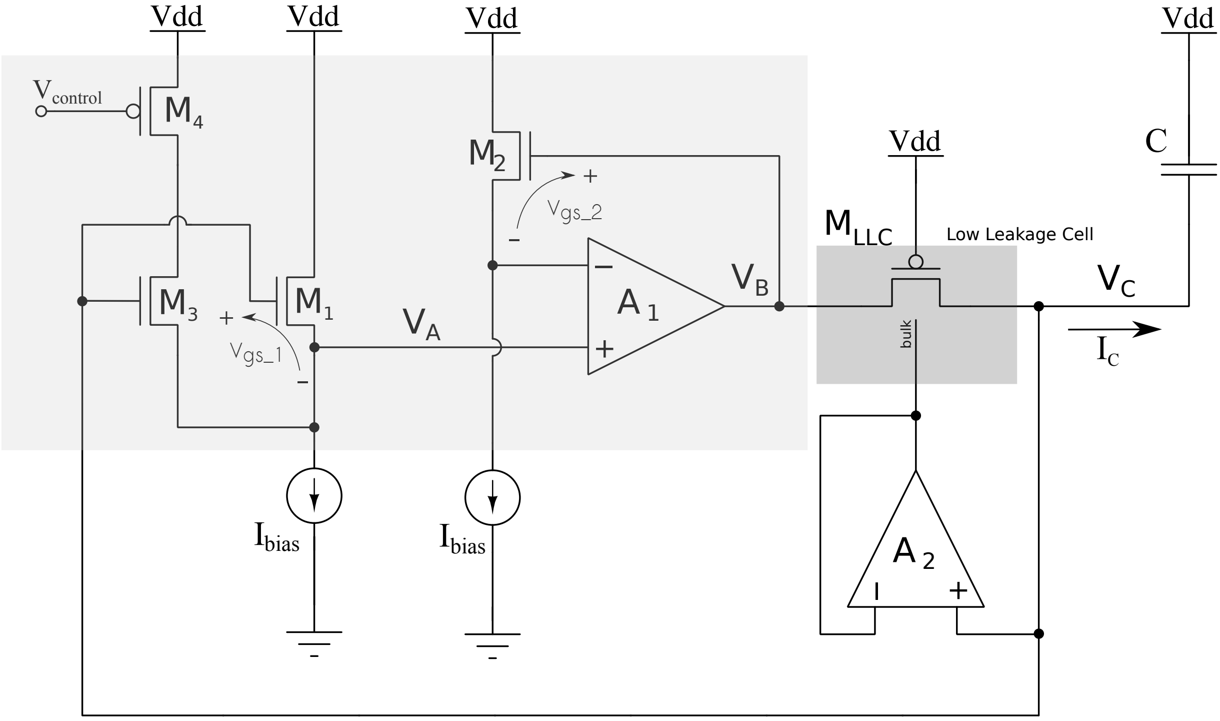

Figure 3.10: Schematic of the Core Idea plus additional circuit that gives the Unbalanced Structure required to get long time constants.

Let's now assume $V_{control} = V_{dd}$. Hence, if $I_{bias} >> I_0$, then the M_4 and M_3 branch is negligible. Thus holds the following KVL:

$$V_B - V_C = V_{gs\_2} - V_{gs\_1} \tag{3.7}$$

Assuming that the I_{bias} current generator are equals and that nMOS sizes are matched is true that:

$$V_B - V_C = V_{gs\_2} - V_{gs\_1} = \frac{U_T}{\kappa} ln\left(\frac{I_{I_{bias}}}{2I_0}\right) - \frac{U_T}{\kappa} ln\left(\frac{I_{I_{bias}}}{I_0}\right) = \frac{U_T}{\kappa} ln\left(\frac{1}{2}\right) \approx -25mV \tag{3.8}$$

Hence, since the I_{db} leakage current is minimized due to A_2, only a tiny current in M_{LLC} flow from V_C to V_B discharging the capacitor C.

Let's now study the $V_{control} = 0V$ case. Now M_4 is above threshold linear region, hence it can be modelled as a low impedance wire. Therefore, M_1 and M_3 are two device in parallel and can be treated as a single device with dimensions $\frac{1\mu m}{1\mu m} + \frac{3\mu m}{1\mu m} = \frac{4\mu m}{1\mu m}$.

Now, inserting this result in Equation 3.7 yields:

$$V_B - V_C = V_{gs\_2} - V_{gs\_1} = \frac{U_T}{\kappa} ln\left(\frac{I_{I_{bias}}}{2I_0}\right) - \frac{U_T}{\kappa} ln\left(\frac{I_{I_{bias}}}{4I_0}\right) = \frac{U_T}{\kappa} ln\left(\frac{2}{1}\right) \approx +25mV \quad (3.9)$$

That means that if $V_{control}$ is low, now the tiny current will flow from V_B to V_C terminals.

I reported here only the analysis of the two extremes values of $V_{control}$. However, there is a certain value $V_{control^*}$ somewhere in the middle of the $V_{control}$ ranges that gives the $V_B = V_C$ condition and hence $I_d = 0$. To calculate the exact value of $V_{control*}$ at which this happens is not straightforward. However, it's not so meaningful since its severely altered by device nonidealities and device mismatch in the AGC loop.

Additionally, the M_{LLC} absolute current is modulated by the value $|V_B - V_C|$. This modulation is hard to predict via simulations at such extreme pMOS polarization conditions. However, as explained later, due to careful design the exact I_{ds} vs $V_B - V_C$ relation won't severely affect the homeostatic circuit behaviour.

Let's now refine Equation 3.7 modelling the current generator mismatch with two current generator with different values: I_{bias} and $I_{bias}+\Delta_{I_{bias}}$, where the term $\Delta_{I_{bias}}$ comprises physical size mismatch and second order effects, such as Early effect. The size mismatch of transisors M_1, M_2 and M_3 can be modelled as a deviation of α from the ideal value and can thus be rewritten as $\alpha + \Delta_\alpha$. Finally, the κ term, in this topology, depends on the the difference of the source terminals of the MOS, i.e. by V_{diff}. Given that, we can hence write:

$$V_B - V_C = \frac{U_T}{\kappa + \Delta_\kappa} ln\left(\frac{\alpha + \Delta_\alpha}{2}\left(1 + \frac{\Delta_{I_{bias}}}{I_{bias}}\right)\right) + V_{diff\_A_1} \quad (3.10)$$

where deviations from ideality of κ results in a increase/decrees of the V_{ds} of LLC pMOS range. On contrary Δ_α, $\Delta_{I_{bias}}$ and $V_{diff\_A_1}$ results in a shift of the V_{ds} range. This last can be a serious issue, especially regarding α and $V_{diff\_A_1}$ terms. In fact, if these two components are remarkable, could happens that the V_{ds} has always the same sign, no matter how is the $V_{control}$ input. Hence, the current in the capacitor can flow only in one direction resulting into a useless circuit.

To be sure that the circuit exhibits both positive and negative V_{ds}, even with such deviations, must be that the V_{ds} range is wide enough to handle worst case scenarios. By simulations $V_{ds} = \pm 25mV$ is a safe value. In addition to that, $V_{control}$ range is expanded by the use of a comparator. This point will be explained later but its purpose is to increase circuit reliability in face of nonidealities.

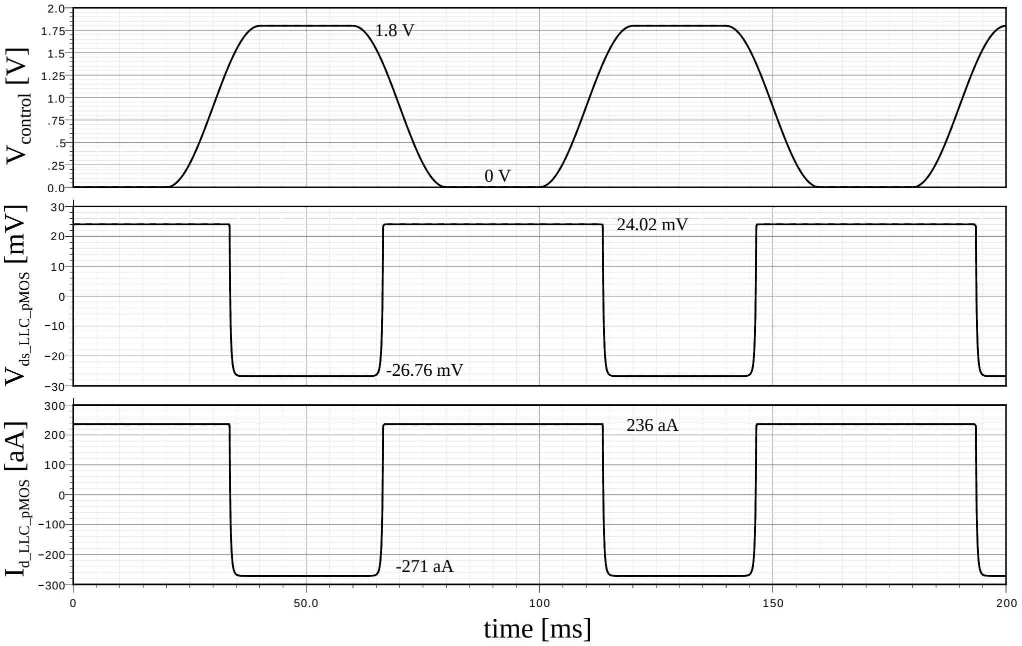

Figure 3.11: A simulation of Figure 3.10 with $I_{bias} = 10nA$, $A = 10000$, LLC $\frac{1\mu m}{1\mu m}$, and $M_4 = \frac{1\mu m}{1\mu m}$. The plot shows how the pMOS I_d current is influenced by the voltage applied to the $V_{control}$ terminal.

In order to unbalance the OPAMP, an alternative approach that I attempted was to add an output or input branch in parallel to the regular one and enable and disable it via a control signal. This technique actually try to affect the V_{offest} of the OPAMP modifying its internal balance and matching of the topology.

Even though several different topologies and combination were examined, they didn't yields linear like behaviours but on contrary a very non regular and ugly piecewise curves. From here the alternative idea to set the V_{ds} of the LLC pMOS not by affecting the inside the OPAMP but by an external circuitry (as was explained in this section). This solution is simulated and reported in Figure 3.11. While $V_{control}$ ranges from ground to V_{dd}, the voltage across the LLC ranges from $24mV$ to $-26mV$ and the current I_C that flow in the capacitor ranges from $236aA$ to $-271aA$.

Furthermore, the external offset technique of this section, in addition to be effective nice-behaviour, is suitable for modular circuit design. In the sense that this concept can be applied even to different types of OPAMP if certain specifications are meet, namely medium gain, very low offset and high input impedance.

This is beneficial in context where OPAMP are already available in design libraries, increasing the flexibility of the structure and facilitate the IC designer.

CHAPTER

4 THE FINAL DESIGN AND SIMULATIONS

In this chapter I close the loop that implements the homeostatic principle in silicon. The loop consist of a dynamic block based on the filter Design 3, a comparator and a differential pair. Simulations that validate the circuits are extensively performed. Details on how to obtain low offset amplifiers are provided at the end of the chapter.

4.1 The Loop Design

Now we have all the elements in order to implement the homeostatic plasticity in silicon. The challenging part of the project was to get ultra long time constants, and it is achieved by on purpose develop the LPF Design 3. The last thing to take into account now is to interface the designed LPF with the synapses and close the AGC loop. However, the Filter Design 3 has voltage input and voltage output, while the artificial synapses have current input and output. Hence, conversion circuits (V-I and I-V) are needed. In addition to that, there are two additional points that I want to illustrate.

First, recalling the first bullet of the *Low Leakage Cell pMOS rule of thumb*, it states that $V_{gs} > 400mV$ must holds always in order to bias the pMOS in accumulation region. But, recalling Figure 3.10 and previous analysis, from the topology point of view there is nothing that assure this condition. Hence, in the loop, an additional circuit must be inserted in order to satisfy this mandatory condition.

The second point results from a more practical consideration. In fact, since we are dealing with 256 neurons in our chip, shared biases among instances is acceptable but is not allowed to have independent biases for tune each of the homoeostatic plasticity circuits that control the neurons. This because it is not feasible in practice due to the inconvenience of setting 256 biases.

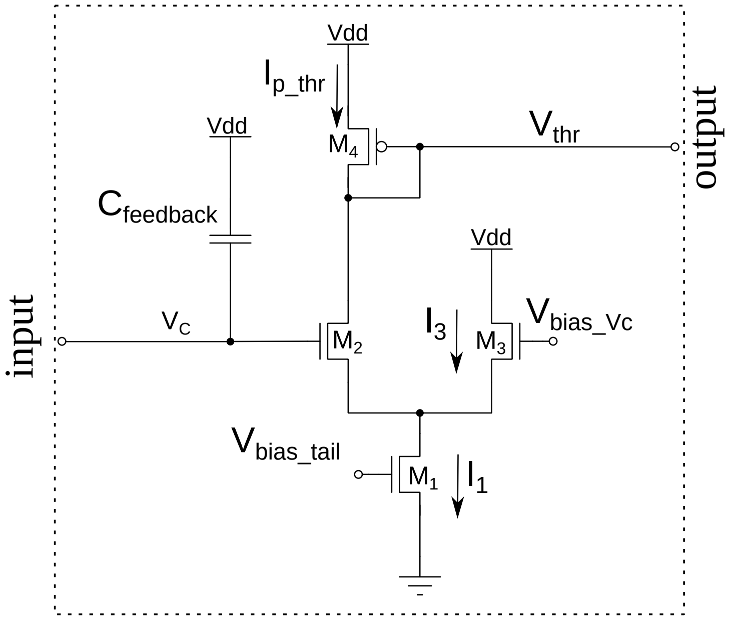

Figure 4.1: The output differential pair circuit. It acts as interface between the LPF with ultra long time constants and the artificial synapse.

Let's now start considering the first point. In Figure 4.1 is depicted a suitable circuit to interface the output of the LPF to the DPI-synapse. It has two purposes, one is to convert the voltage output of the LPF (V_C, the voltage across the capacitor) into a current I_{p_thr} suitable to be used as gain control signal for the DPI. The second is to set the DC value of V_C in order to bias the pMOS in accumulation mode.

However, I_{p_thr} value is set by the feedback loop itself and can't be directly modified. In particular, the dynamic of this current is associated to the V_C dynamic that can't be modified either.

Let's now analyse the differential pair of Figure 4.1. If M_1 is in subthreshold saturation region, Equation 2.19 holds and yields:

$$I_1 = I_0 \frac{W}{L} e^{\kappa \frac{(V_{bias\_tail} - V_s)}{U_T}} \tag{4.1}$$

the same is true for nMOS M_2 and M_3 giving:

$$I_{p\_thr} = I_0 \frac{W}{L} e^{\kappa \frac{(V_C - V_s)}{U_T}} \tag{4.2}$$

$$I_3 = I_0 \frac{W}{L} e^{\kappa \frac{(V_{bias\_Vc} - V_s)}{U_T}} \tag{4.3}$$

applying the KCL at the source terminal of M_2 and M_3 is true that:

$$I_1 = I_{p\_thr} + I_3 = I_0 \frac{W}{L} e^{\frac{-V_s}{U_T}} \left(e^{\frac{\kappa V_C}{U_T}} - e^{\frac{\kappa V_{bias\_Vc}}{U_T}} \right) \tag{4.4}$$

and solving for V_s and substituting in Equation 4.2 of I_{p_thr} finally yields:

$$I_{p\_thr} = I_1 \frac{e^{\frac{\kappa V_C}{U_T}}}{e^{\frac{\kappa V_C}{U_T}} - e^{\frac{\kappa V_{bias\_Vc}}{U_T}}} \tag{4.5}$$

So, since I_1 and V_{bias_Vc} are set by biasing the circuit and I_{p_thr} is set by the feedback loop, the DC value of V_C is hence a dependent variable. Such a value can be chosen by properly size the MOS and set the biases. In fact, from Equations 4.2 and 4.3 $V_C = V_{bias\_Vc}$ if $I_{p\_thr} = I_3$ and the size of M_2 is equal to the size of M_3. Current I_{p_thr} and hence I_3 can be estimated by Equation 2.37 once given the biases of the DPI- synapses.

In normal operation condition I_{in} changes, and results in a I_{p_thr} variation and thus in a V_C variation that counterbalance the DPI input changes effect. Quantitatively, V_C AC signals are usually (simulated) small (up to $\pm 100mV$) compared to its DC value. Its swing ΔV_C is inversely proportional to the gain T of the loop.

Let's now consider the LPF input stage voltage swing. The control signal $V_{control}$, has a full range input swing that goes from ground to Vdd. Inside this range there is a particular value $V_{control*}$, dependent on the MOS size, biases and on process variations, that gives a on-set filter condition, i.e. $V_B - V_C = 0 \rightarrow I_C = 0$. In order to let the AGC circuit to work properly, the input signal of the LPF must has a swing that can crosses $V_{control*}$. This is the condition that allows the filter capacitor to charge and discharge giving ultra long temporal dynamics. See Figure 4.4 for a visual representation of the running signals. **(a)** represent a working condition since $V_{\epsilon-} < V_{control*} < V_{\epsilon+}$, while **(b)** doesn't because $V_{control*}$ is not in the reachable range of V_ϵ.

However, if we implement the summer (the block that drives the LPF) as in Figure 4.2 with a simple three MOS structure, its output voltage V_ϵ swing is very limited in amplitude. As emphasized before, this potentially can results in the impossibility to properly drive the LPF if V_ϵ can't crosses $V_{control*}$. Even though the output DC voltage of the summer can be set by proper sizing the transistors of the summer, due to very small summer output V_ϵ voltage ranges (mV), mismatch and variation among instances can still shift the range of V_ϵ away from $V_{control*}$.

A solution to that is to add an amplifier (CCVS) in cascade of the summer in order to increase the range of the summer and let it to cross $V_{control*}$, Figure 4.3 **(b)**.

We actually implemented this solution in a very conservative way by replacing the amplifier (CCVS) with a comparator (that can be thought as an amplifier with really high gain and saturation limits). Figure 4.3 **(c)**. This comparator has the full power supply output range V_ϵ that allow us to be sure the circuits is able to properly work even without additional bias. However, this architecture pays in terms of reliability of the circuit but introduces an unwanted dependence of the time constant of I_{syn} to the input signal that will be addressed later.

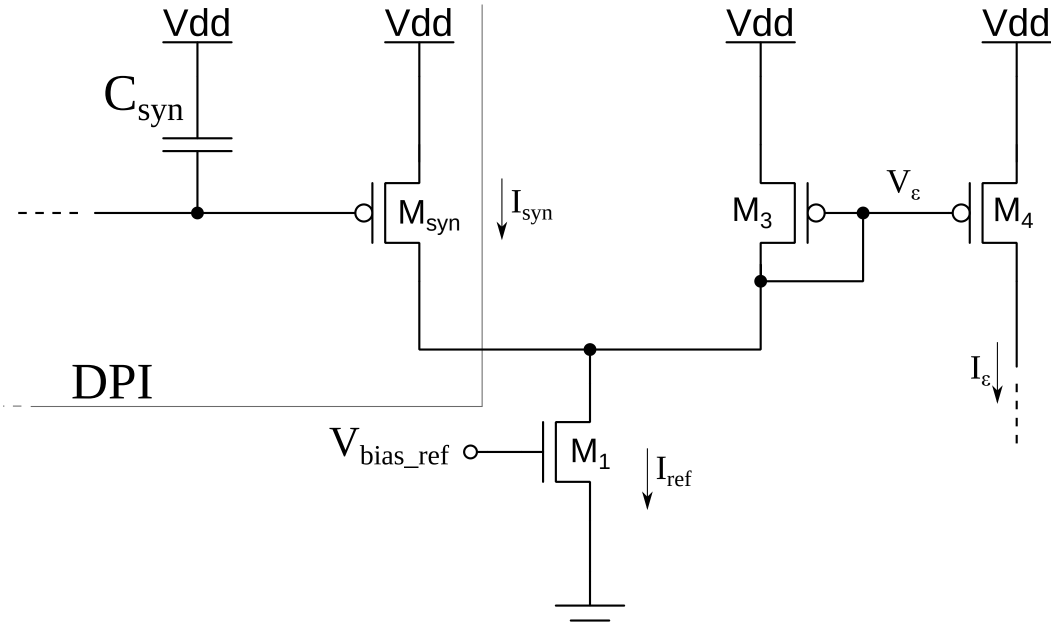

Figure 4.2: A simple three MOS summer exploit KCL. The output range of V_ϵ doesn't extend from ground to V_{dd}.

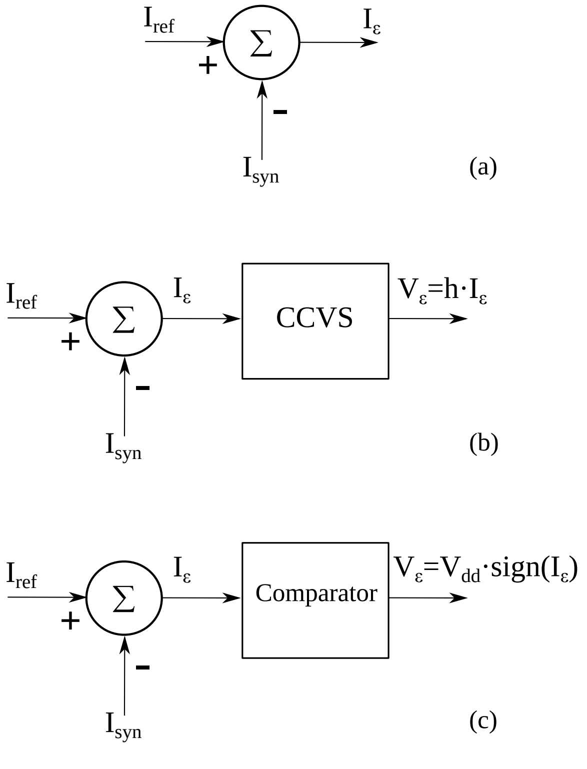

Figure 4.3: Schematic of the usual feedback summer (a), the same summer with an amplifier in order to increase the output range (b) and the actual implemented version (c) that is a degeneration of solution (b). Solutions (a) and (b) are linear while solution (c) is hihlgy non linear.

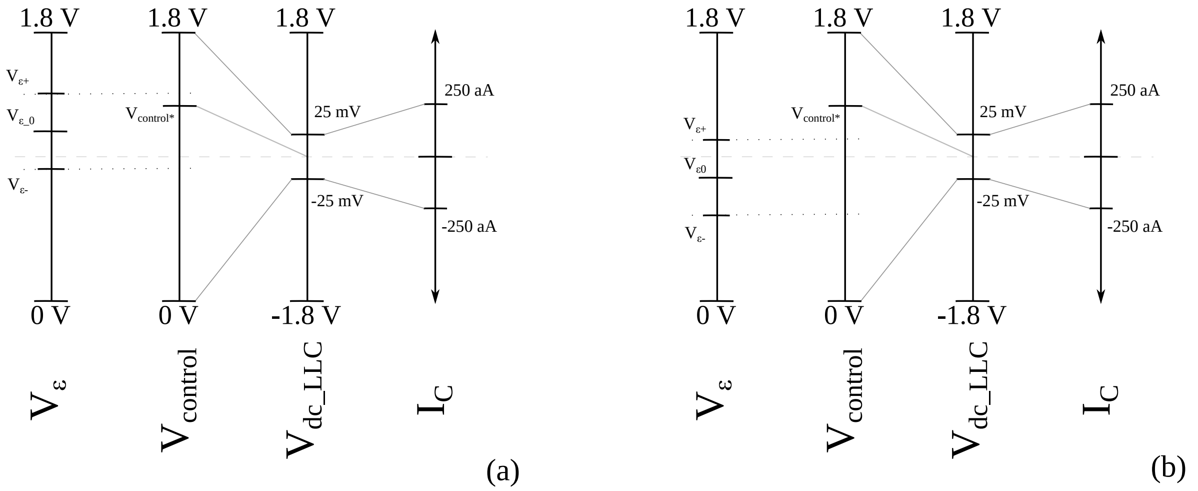

Figure 4.4: Visual representation of the signal ranges in the loop chain. This representation gives a visual understanding of the problem of limited range of V_ϵ. (a) is the case in which the signal V_ϵ range includes $V_{control*}$ resulting in a proper working circuit. (b) On contrary represent a cicuit that can't work because V_ϵ doesn't include $V_{control*}$ in its range, hence $V_{ds\_LLC} < 0$ and $I_C < 0$.

$V_{bias\_DPI\_weight}$	I_τ	V_{bias_WTA}	I_{ref}	I_{bias}	V_{bias_tail}	V_{bias_Vc}	$V_{start-up}$
$1.8V$	$4.5pA$	$335mV$	$10nA$	$10nA$	$230mV$	$1.2V$	0 then Vdd

Table 4.1: Biases of schematic Figure 4.5.

The implemented comparator is just the modified WTA circuit described in Section 2.5. It acts as a comparator with input current signals and output voltages. The whole designed circuit with all blocks is shown in Figure 4.5.

Since we are dealing with very small currents few additional remarks about simulations are needed. The first is that, in order to get meaningful results, default simulation parameters are often not adequate. In our case, with Cadence ADE, I set $gmin = 10^{-14}$ and $iabstol = 2 \times 10^{-17}$. These are the lowest value that provides nice results and the convergence on the simulation.

MOS	W [μm]	L [μm]	MOS	W [μm]	L [μm]	MOS	W [μm]	L [μm]
M_1	1	1	M_8	10	0.5	M_{15}	1	0.5
M_2	2	1	M_9	1	0.5	M_{16}	1	0.5
M_3	3	1	M_{10}	1	0.5	M_{17}	1	0.5
M_4	1	1	M_{11}	1	0.5	M_{18}	1	1
M_5	4	0.5	M_{12}	1	0.5	M_{19}	1	1
M_6	4	0.5	M_{13}	1	0.5	M_{20}	1	1
M_7	2	1	M_{14}	1	0.5	M_{21}	2	1

Table 4.2: The list of the actual MOS sizes for the design of Figure 4.5.

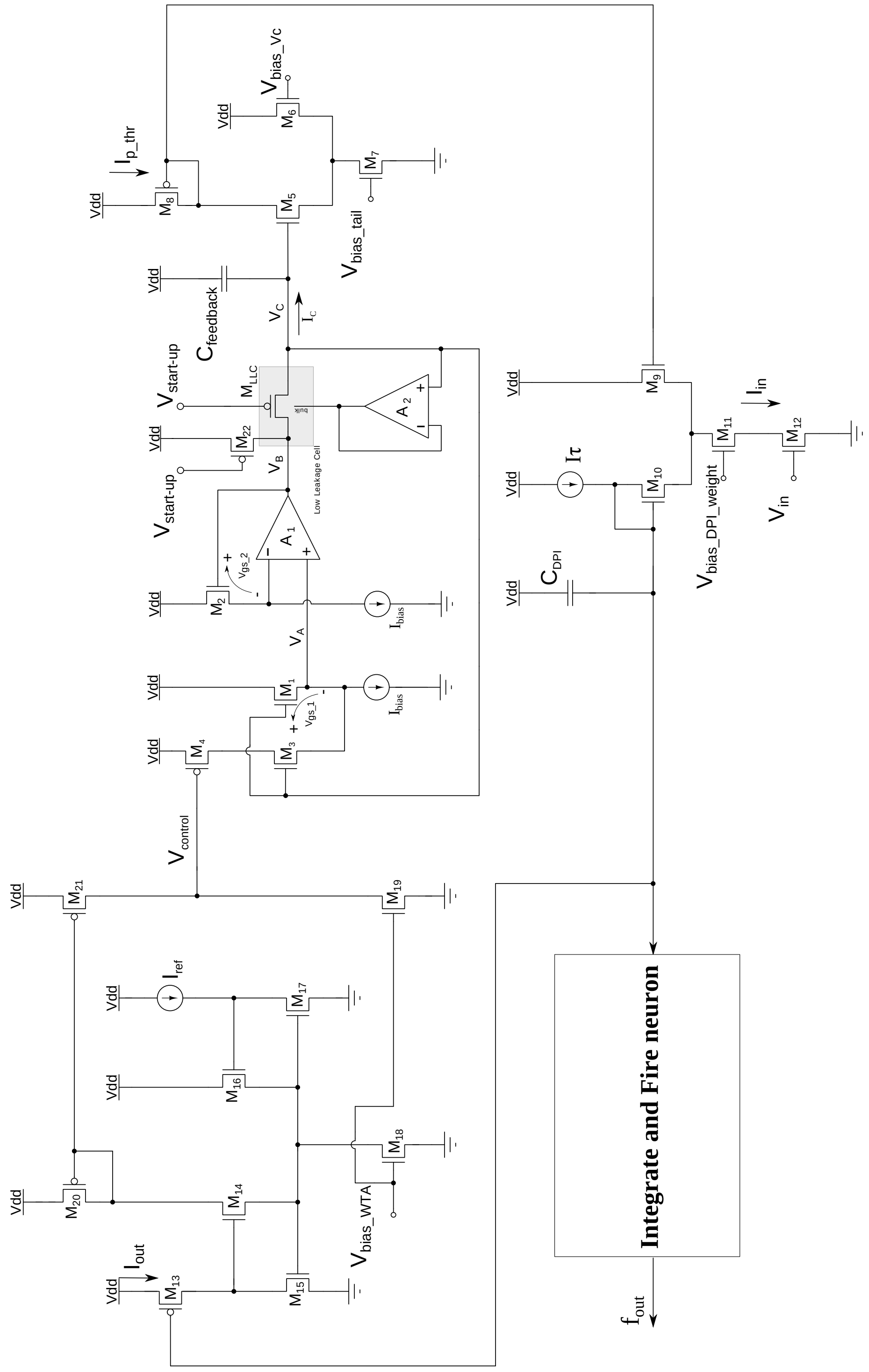

Figure 4.5: The final schematic of the homeostasis AGC in silicon. It comprises the WTA, the LPF and the output differential pair. At the very beginning, as first thing $V_{start-up}$ must be set to $0V$ in order to let be $V_C = 1.8$. This is the start up condition, then $V_{start-up}$ must be set equal to V_{dd}. This lets the circuit works as explained in the text with long time constants.

The second remark is that low currents gives long time constants in the homeostatic circuit loop. These dynamics are order of magnitude higher that DPI dynamics hence, simulation of these two systems is usually very resource consuming resulting long lasting simulations. This happens because simulation must perform, for long time windows (due to LPF long time constants), short time step analysis (due to the fast dynamics of the DPI). However, since the two dynamic magnitudes are so different, the DPI I_{syn} fast dynamic can be approximated with its mean value. Hence, instead of feed the input of the DPI with a fast digital input and obtain the relative fast spike out I_{syn} of the DPI, I stimulated the DPI with current mean values. I.e. I_{in} represent the mean value input current of synapses that stimulate one neuron.

A simulation of the final design sized as Table 4.2 and biased as Table 4.1 is shown in Figure 4.6. The input signal is provided to the synaptic input I_{in} of the DPI. It is a current square wave that represent the mean current of all the 256 synapses that are connected to one neuron. Hence, increase in I_{in} means that synapses are stimulating the neuron more intensively. I_{in} varies from around $4nA$ to $22nA$ while the homeostatic reference, i.e. the synaptic current at which the system would converge, is $I_{ref} = 10nA$. Without the homeostatic circuit these condition would results in $I_{syn} \neq I_{ref}$ that lasts for ever. But, with the homeostatic circuit, the system I_{syn} will converge to I_{ref} with long dynamics. This is exactly what can be observed from simulation of Figure 4.6. In the second window is plotted I_{syn} and, after a I_{in} changes, I_{syn} suddenly changes as well because the gain of the DPI is still the same. But, after a certain amount of time, that is proportional to C and I_C, I_{syn} is forced to converge at $I_{ref} = 10nA$ due the homeostatic effect. From the quantitatively point of view we see that this transitions lasts $260s$ or $240s$ according to the sign of I_C in the capacitor, that is not equal in the charge and discharge phases. In fact, since I_C is proportional to $\Delta V_{ds\_LLC\_pMOS}$, observing window 3 of Figure 4.7, we see that $\Delta V_{ds\_LLC\_pMOS}$ in the charging phase is $-1.6 + 24.5 = 22.9mV$ and yields a $\Delta t = 260s$. On contrary in the discharging phase, $\Delta V_{ds\_LLC\_pMOS} = 22.7 + 1.6 = 24.3mV$ that gives $\Delta t = 240s$. The value of $-1.6mV$ is the steady state condition that compensate leakages currents (most likely due to the A_1 and A_2 offsets and leakages at gates) in order to maintain $I_C = 0$.

Deviation of $I_{syn} = 10.54nA$ at steady state from $I_{ref} = 10nA$ are due to nonidealities in the comparator block (the WTA). This is not a serious issue since gives only an offset to I_{syn} that can easily fixed by changing I_{ref}, if needed.

The shape of the I_{syn} is not exponential as happens in any first order linear system. In fact, the presence of the comparator in the loop gives a highly non linear behaviour that is responsible for this response of I_{syn}. This happens because with the comparator we loose the information how close is I_{syn} compared to I_{ref} hence, the capacitor is always charged or discharged with constant currents. This can be observed in Figure 4.7 (central window) that shows $V_{control}$, i.e. the output of the WTA comparator, which value can be either high ($1.799V$), low ($5\mu V$) or somewhere in the middle $V_{control*} = 1.395V$. There are no in between values that prove that the circuit can process only the sign of the difference of $I_{syn} - I_{ref}$, but not its magnitude. Finally, simulation of Figure 4.8 plot the V_C value in response to the usual square wave current input. The V_C DC voltage is around $0.995V$ with an AC signal $\pm 30mV$. This condition assures both non distortion in the output differential pair of Figure 4.1 and the accumulation mode bias of the LLC pMOS ($V_{gs} = 1.8 - V_C \approx 0.8 > 0.4$).

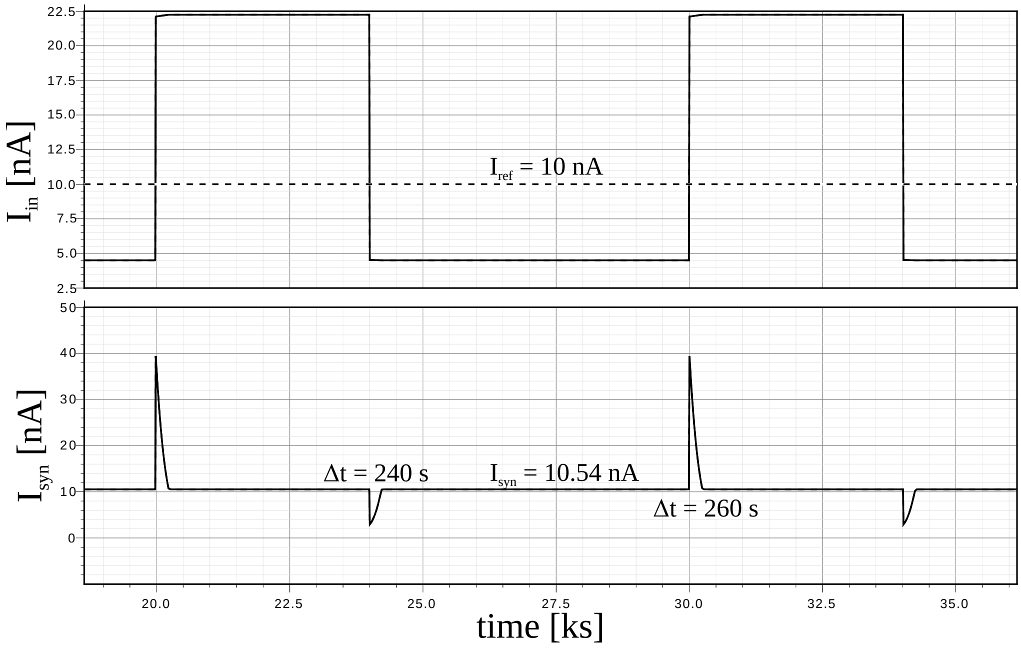

Figure 4.6: Simulation of the Final Design of Figure 4.5. The top window represents the input signal $I_{in} = 4nA \div 22nA$ feed into the DPI. The reference for the homeostatic plasticity is $I_{ref} = 10nA$. The lower window plots the DPI output synaptic current, its dynamics are around 4 minutes with 1pF capacitor.

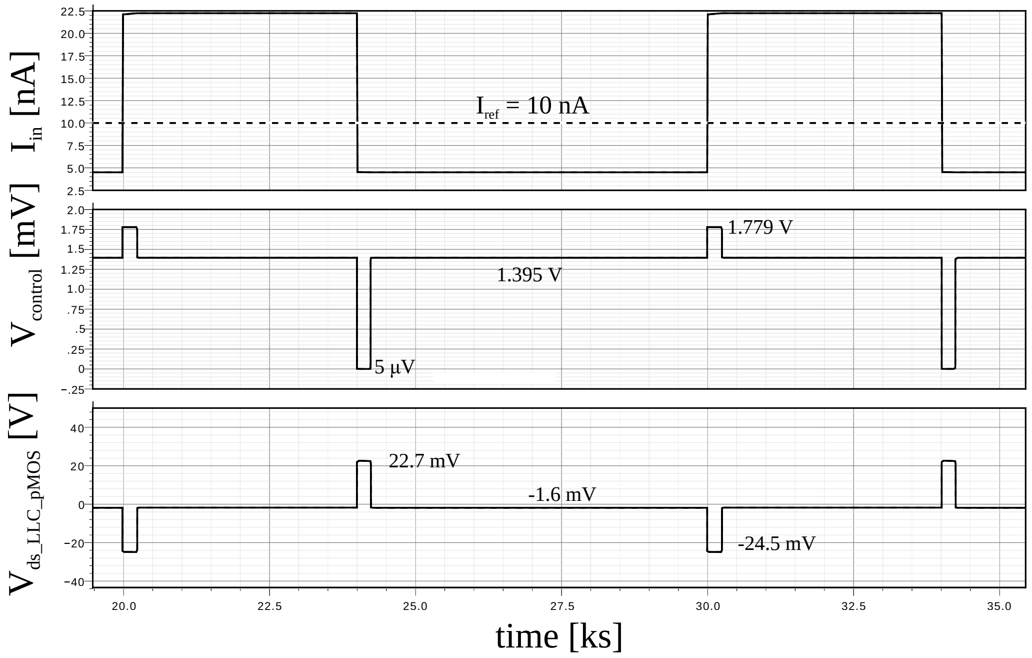

Figure 4.7: This simulation shows $V_{control}$, (i.e. the output of the WTA comparator) and the voltage difference across the LLC pMOS in response to the I_{in} stimuli.

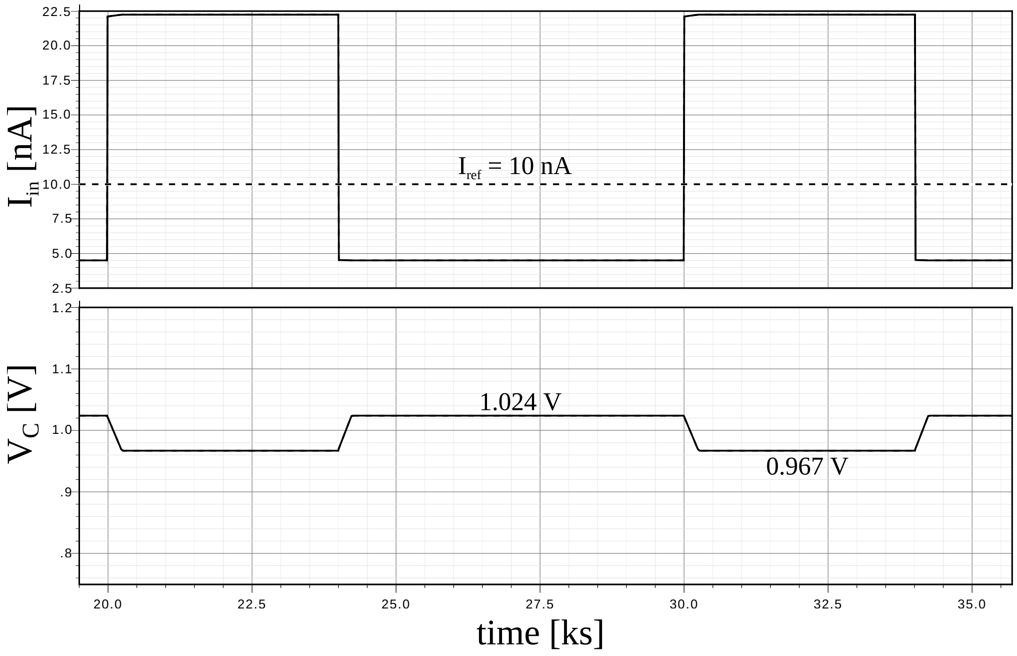

Figure 4.8: This simulation shows the voltage in the state variable capacitor. Its value is always well under $1.4V$ in order to satisfy the accumulation mode condition for the LLC pMOS ($V_{gs} > 400mV$).

Previously, we mention that the time constant of the circuit is strongly dependent to the value of I_{in}. In order to understand this behaviour is important to note that the time constant by which the capacitor voltage is modified is directly proportional to C and to $\frac{1}{I_C}$. However, the loop gain T is proportional to the input current I_{in} of the DPI, see Figure 3.2. The higher is the gain, the less are the AC variation of V_C that changes I_{p_thr} and hence the gain of the DPI synapse. Vice versa if the loop gain is small, it results in high variations of V_C in order to counterbalance the effect of I_{in} variation.

This is confirmed by circuit simulations of Figure 4.9. In fact, with $I_{in1} = 1n \div 6nA$ the loop gain T_1 yields $\Delta_{V_C} = 64mV$, with $I_{in2} = 4n \div 22nA$ the loop gain T_2 yields $\Delta_{V_C} = 57mV$ and with $I_{in2} = 11n \div 44nA$ the loop gain T_3 yields $\Delta_{V_C} = 54mV$. Is clear that the more is the input I_{in} the higher is the gain T and the lower is Δ_{V_C}. However, the slope by which V_C varies is constant and proportional to C, $\frac{1}{I_C}$.

Given that, is clear that the effective time by which the system will get to the equilibrium state is directly proportional to C, $\frac{1}{I_C}$ and ΔV_C, that is the above mentioned difference between the voltage on the capacitor taken at the steady state points with two different inputs I_{in}.

This variation in dynamics is a problem that didn't affect the operation of our specific application, but could be relevant in other contexts. This issue primary results from the use of a comparator instead of a linear summer, even though this would not be sufficient either to obtain exponential time dynamics [16]. In fact, by using the comparator we loose the information about how far is I_{syn} from I_{ref}. Hence, from the above considerations, the V_C slope and the voltage difference of V_C related to two equilibrium points both contributes to the effective time constants.

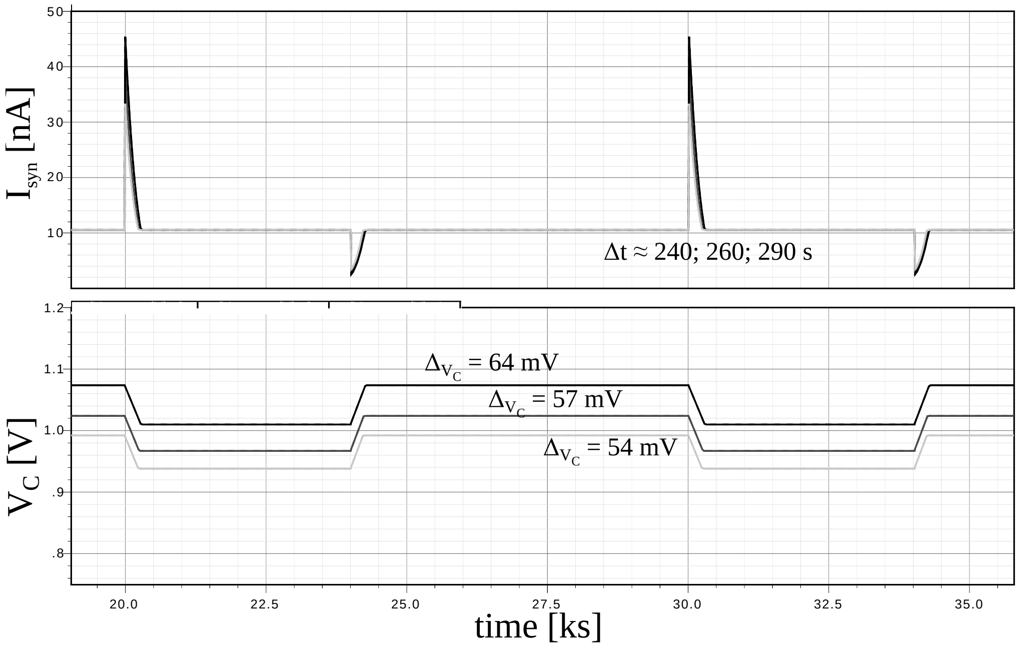

Figure 4.9: This simulation of the loop shows how I_{syn} dynamics varies according to Δ_{V_C}. Δ_{V_C} is inversely proportional to the loop gain T set by changing I_{in}.

4.2 A low offset amplifier

In this section I finally focus on how to design simple low offset amplifiers that are needed in my design. The idea described in this section is presented in [14] and can be applied only to a unity gain amplifier configuration.

As contrary as the auto zero offset cancellation technique, this approach doesn't use any continuous auto calibration but instead rely on matching properties. Hence, the design results in a very compact and effective topology that nicely fits in my LPF.

Usually, the offset issue is an unwanted deviation from the ideal amplifier model that mainly results from mismatch considerations and from second order effects in MOS devices. It is defined as the "differential voltage that has to be applied to a differential amplifier in order to cancel the DC offset output voltage". In a simple 5 MOS OTA of Figure 4.10 (tail generator plus differential pair plus current mirror) happens that, with $V_{in+} = V_{in-}$ with proper V_{in_common}, the currents in the two differential pair branches (VINN, VINP) are different if transistors are not matched and if V_d voltages at drains of paired nMOS are not equal. This last point happens systematically, no matter how well matched are MOS in the layout, because the OTA structure has not a perfectly symmetry topology. In fact, M_4 pMOS is diode connected ($V_d = V_g$) while the respective MOS in the other branch M_5 is not. Hence, even with the same V_{gs} applied, currents of M_4 and M_5 in the mirror are not perfectly copied due to channel length modulation. Given that, a simple way to reduce the offset effect is to make transistors bigger and to add MOS cascode in order to reduce this channel modulation issue.

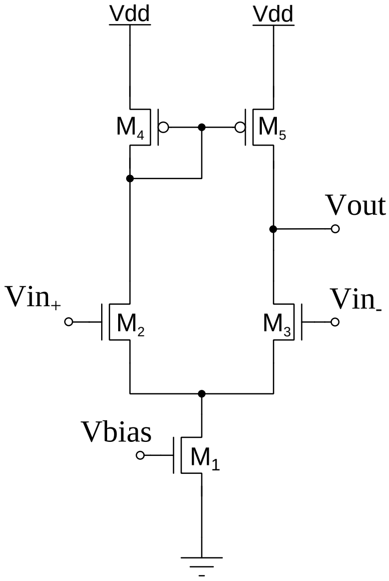

Figure 4.10: A basic five MOS Operational Transconductance Amplifier (OTA).

However, even careful designs, with such technique offset only down to $5mV$ can be obtained. This has been simulated and is shown in Figure 4.11, which plot has four windows. The first one is the sweep of the differential input voltage of the OTA, the other three windows are the output voltages of the OTA.

The simulation shows that, even though every transistor in the schematic has perfect sizes, there is still an offset of $0.9mV$ resulting from the non symmetry of voltages in the two OTA branches. However, this is not even a realistic case. In fact, in real circuits, meaningful deviations between transistors ratio are experienced.

In Figure 4.11, I simulated the offset voltage of the OTA by changing of 20% one MOS size, namely the differential pair pMOS (M_3) or the current mirror nMOS (M_5). This unpaired situation gives different currents in M_3 and M_5 that results in $V_{out} \neq V_{common}$ and hence in a output offset.

From simulations of Figure 4.11 is clear that an offset around mV is something that can easily happens in real circuit implementations. This is not tolerated in our application since we are dealing with voltages across the Low Leakage Cell in the order of $50mV$.

The solution exploited by [14] aims to make zero offset by matching two OTA offsets and subtract each others. In their paper the authors claimed that the obtainable offsets are in the order of μV (simulated). The underlying idea is depicted in Figure 4.12 and shows a main amplifier A_{main} that has some offset V_{offset_main}. Hence, in a voltage buffer configuration holds that $V_{out} = V_{in} + V_{offset\_main}$.

So, if in the unity gain loop path there is a DC voltage source with the same value V_{offset_main}, and if we are able to somehow subtract it from V_{out}, then the overall offset would be instantaneously cancelled and $V_{out} = V_{in}$, resulting in an offset free signal. But, since V_{offset_main} is not known and is variable, the best way to implement that voltage generator

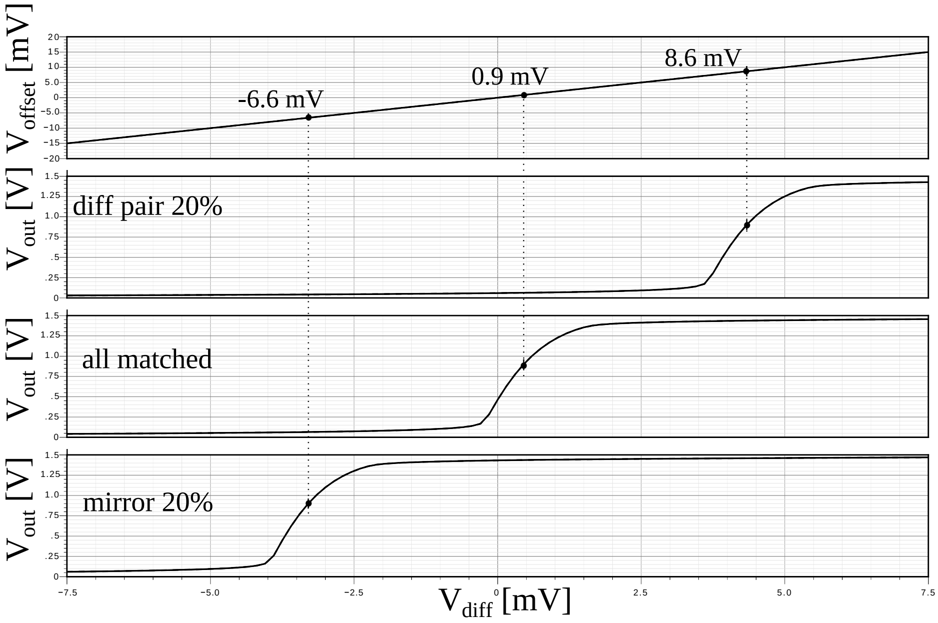

Figure 4.11: Voffset simulation of a five transistor OTA with cascode. The circuit is loaded with a $10pF$ capacitor and biased with $I_{tail} = 200nA$. There simulations are plot, one with all transistors matched, the other two with a deviation of 20% from the nomianl size

is to have a copy of A_{main} called A_{fb} that provides a good estimation of V_{offset_main}.

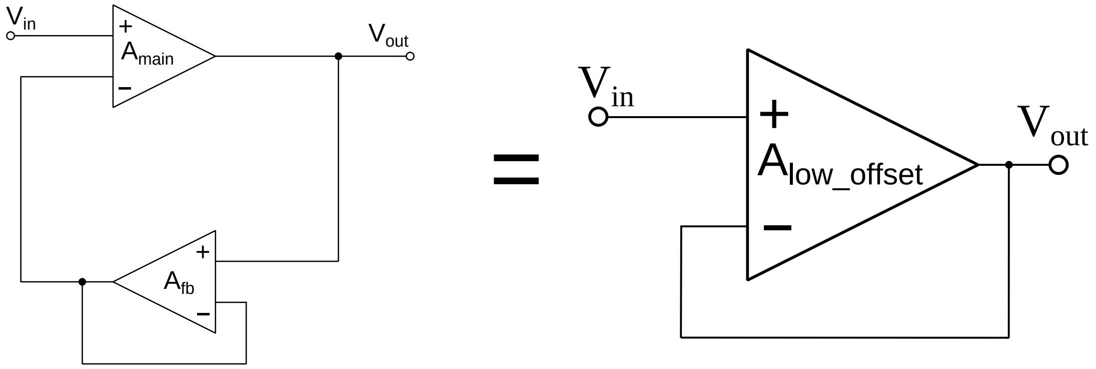

Figure 4.12: A technique to reduce offset is to place two matched amplifier in a loop. This arrangement cancels the offsets of the two devices in the signal path resulting in a effective low offset unity gain amplifier.

The CMOS implementation of amplifiers of Figure 4.12 is depicted in Figure 4.13. It consist of two simple five MOS OTAs with cascodes (M_2, M_6, M_{12}, M_{13}, M_9, M_5 and M_3, M_7, M_{10}, M_{11}, M_8, M_4) and with one shared tail generator M_1.

Matching considerations suggest to match same transistors of the two different amplifiers rather than matches transistors in the same amplifier (as would have be done in regular design). For example, M_2 should match M_3 by being placed close in the physical layout, M_4

and M_5 as well, and so on. Applying this criteria it doesn't minimize the offset of the single amplifier A_{main} nor the one of A_{fb}, but, on contrary, it matches the magnitude of the two offsets of the two amplifiers. Then, since they are subtracted each other by the topology in the loop path, this strategy minimizes this difference between the input V_{in} and the output V_{out} of A_{low_offset} (see Figure 4.12) with matched non minimum A_{main} and A_{fb} offsets.

To further reduce the sources of mismatches, the bias of the amplifiers is shared and provided by M_1. Hence, OTAs are no longer independent but, since gate and source voltages of MOS M_2, M_3 and M_4 , M_5 are ideally identical, the effect on the circuit behaviour is small [14]. As benefit, this topology forces source voltages of those pMOS to be identical enhancing the matching characteristics between the two amplifiers.

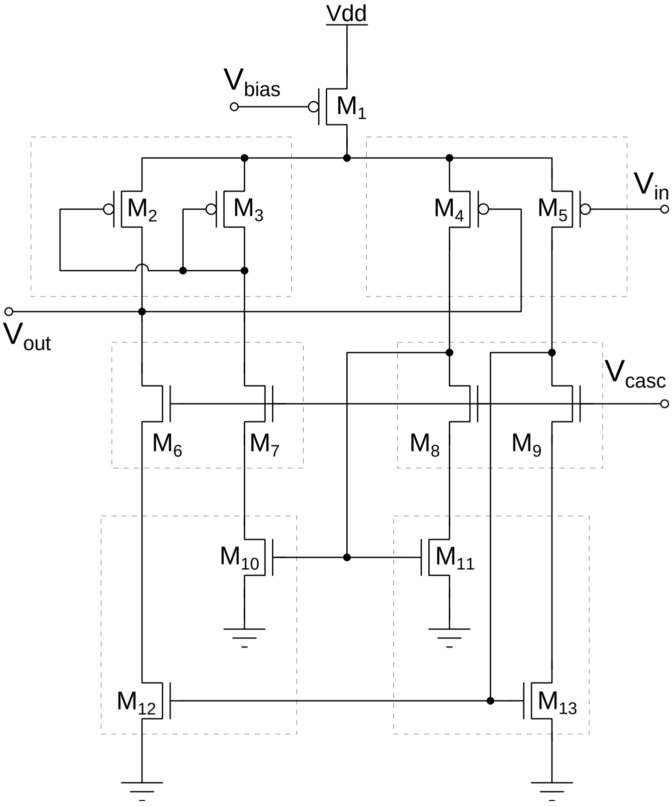

Figure 4.13: The transistor level implementation of A_{low_offset}. M_2, M_6, M_{12}, M_{13}, M_9 and M_5 realizes the A_{main} amplifier while M_3, M_7, M_{10}, M_{11}, M_8 and M_4 realizes the A_{fb} amplifier. The current source M_1 is shared between the two amplifiers. Matched MOS are grouped by dotted boxes.

In Figure 4.14 there is plot the simulation of schematic of Figure 4.13 and shows how the structure of Figure 4.12 works well and has good offset performances. The input signal is a DC value at $0.9V$ with superimposed an AC $\pm 10mV$, 100 Hz sinusoid. The first window in the plot is the difference between V_{in} and V_{out} of the amplifier. In the first window there are two curves, one relative to all matched MOS and the other relative to a 20% variation of transistor M_4 and M_5.

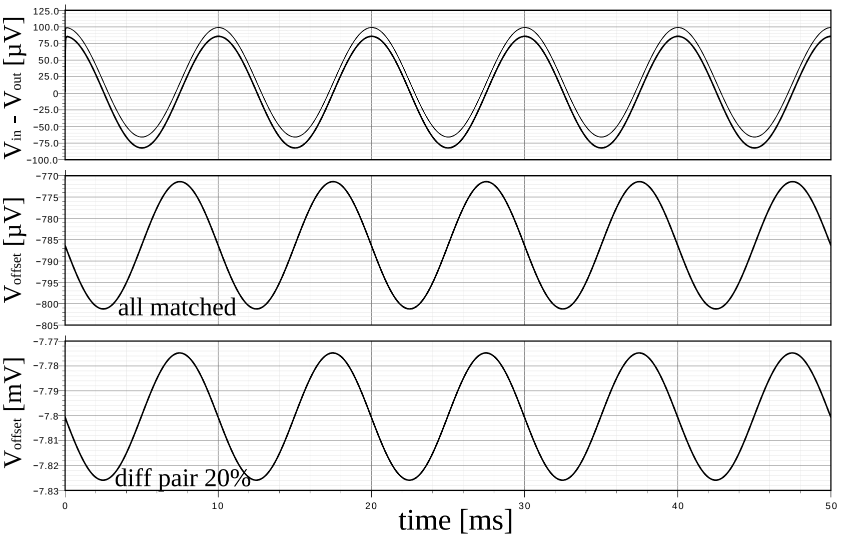

Figure 4.14: The first window shows the difference between the input V_{in} and the output V_{out} of the amplifier, one with all transistor matched, and the other with M_4 and M_5 size increased by 20%. The difference $V_{in} - V_{out}$ is almost the same between the two curves, even though the amplifiers offsets are way different. Window 2 and 3 respectively shows the V_{offset} of A_{main} in the two matching conditions.

CHAPTER 5 CONCLUSIONS

5.1 Discussion

The homoesostatic principle is a key mechanisms that allows biological neurons in large neural network to interact each other and work properly even with high variations of chemical concentrations and physical quantities. Due to its effectiveness, a good idea is to transpose this concept even in artificial neurons populations. The most straightforward implementation of the homeostatic plasticity in silicon is to build an Automatic Gain Control loop around the artificial neuron. This allows to change the neuron input-output gain by modifying its synaptic gain and thus keep the output firing rate of the neuron at a reference frequency. The homeostatic plasticity is very useful in order to face process variation mismatches, temperature variation, changes in chip loads and so on. The homeostatic mechanisms can be combined with Hebbian synaptic plasticity on the same neuron in order to provide a wide range of adapting mechanisms. In fact, has been proven that complex behaviour are obtained by the interaction between homeostatic and Hebbian plasticity that can't be reproduced by one single mechanisms alone.

The key blocks of my AGC loop homeostatic plasticity implementation are a LPF, a comparator and a differential pair. A very important specification of the AGC loop is that it requires ultra long dynamics but still must exhibits compact size due to chip integration reasons. From these two specifications result opposite design styles that usually yields in trade-offs.

Due to its appeal, some attempts in past literature has been performed in order to implement this homeostatic plasticity in artificial neural networks. Unfortunately, they required floating gate design technique that needs higher voltages and higher area. Other designs exploit workstations to generate long time constants, preventing it from the use in portable applications.

An idea for obtain ultra long time constants, while still occupy small area, is to deal with tiny currents in the order of magnitude of atto-Ampere. Such small currents can be generated by properly biasing a pMOS in accumulation mode and reducing all leakage mechanisms. From these insights, I first I developed a femto/atto ampere current generator (Section 3.2), then I developed a novel unbalanced architecture (Section 3.5) based on it that can charge/discharge the system state capacitor with such tiny currents and thus obtain ultra long time constants.

In this thesis work I first understand and analysed the homeostatic principle in neurons by reading the current literature. Then I model it and reformulate the problem in an engineering way with an AGC. The solution I proposed here in my thesis work meet the specification of compactness and ultra long dynamics described in Section 1.5. Results are validated by extensive software simulations (Cadence ADE) and plots are reported in Figure 4.6, Figure 4.7, Figure 4.8 and Figure 4.9.

Here I want to mention that I'm aware that I'm simulating the pMOS (Low Leakage Cell) in odd biasing condition where usually, in standard analog design, no high precision is required. As far as I understand, the pMOS model (PSP) is quite realistic, but the actual overlapping behaviour between the Cadence simulated pMOS and the real fabricated pMOS strongly depends on the foundry device characterization that provides parameters for the Cadence PSP model. Hence, if a not accurate real pMOS characterization were performed by AMS, the simulation results are quantitatively imprecise.

However, I'm optimistic about the success and feature of the circuit because the designed structure intuitively makes sense and it is supported by the qualitatively analysis. Even though simulations can gives not accurate results, since my structure is based on [6] and [7] where true on chip measurements were performed, the order of magnitude of the running currents must be comparable.

Due to these simulation limitations, true on chip measurements must be performed in order to validate and full characterize the designed structure. This will be done when the chip will be sent back from the foundry (Fall 2013).

As a conclusion I briefly provide here the features and limitations of my implementation of the homeostatic circuit in silicon. It satisfies all the specifications of the project (see 1.5) while introducing not severe constraints (for our application).

Pros

- Ultra long dynamics $\Delta t \approx 250s$ with realistic biases and signals (simulated up to 600s).

- Very compact design 16 x 46 μm (LPF, amplifiers, comparator, differential pair).

- No calibration is required and mismatch robustness.

- Power consumption of $100nW$ (comparator, LPF, amplifiers, output differential pair).

Cons

- Simulations results too dependent on software parameters, most likely due to the inaccurate model fit of pMOS in accumulation region.
- Non-linear dynamics due the comparator.
- Time constant dependent to the input I_{in} magnitude.

5.2 Future Works

Due to short time available for develop the full working circuit, surely its design is not optimized for each part and then could be further enhanced. If I'll have the opportunity, in the future I would like to spend time to make it better, especially in terms of time constants and linearity. In fact, since my design was quite conservative, I think that even better results in terms of filter cut-off frequency can be obtained without sacrifice robustness and reliability. In addition to that, another good improvement for the homeosatatic loop would be to substitute the non linear comparator with a pure analog summer in order to have linear dynamics.

If the circuit will prove to work in real chip, this unbalanced filter technique can be effectively used where ultra low cut-off frequencies are required, such as pace-makers, averaging, real world signal conditioning interfaces and so on [20]. Here below I report in Table 5.1 the state of the art sub-Hertz LPF comparison, where the Normalized Cut-off Frequency is obtained by multiplying the Cut-off Frequency by $1pF/Integrating\_Capacitance$. This parameter allow a fair comparison between the designs cut-off frequencies as if they had the same capacitor. For my work, the cut-off frequency is obtained by treating dynamics as if they were exponential in a linear system. Hence, according to simulations, $\Delta t \approx 250s$ (average) and assuming that the dynamics transients ends in 5τ, in my design $\tau = 250/5 = 50s$, that gives a $f_c = \frac{1}{2\pi\tau} = 3mHz$.

Reference	CMOS process [μm]	Cut-off Freq. [Hz]	Supply Voltage [V]	Area [mm^2]	Integrating Capacitance[F]	Normalized Cut-off Freq. (1pF) [Hz]
[9]	0.35	0.5	-	-	100f	0.035
[21]	0.35	35	3.2	0.025	25f	0.875
[22]	0.50	0.180	-	0.035	15p	2.7
[23]	1	0.075	5	0.25	10p	0.75
This work	0.18	0.003	1.8	0.0012	1pF	0.003

Table 5.1: A state of the art comparison of sub-Hertz flters. Table partailly taken from [20].

Table 5.1 shows very good performances in terms of cut-off frequencies for my design. In particular, my Normalized cut-off frequency is one order of magnitude lower than state of the art design but with far less required area.

APPENDIX

A

MASKS LAYOUT

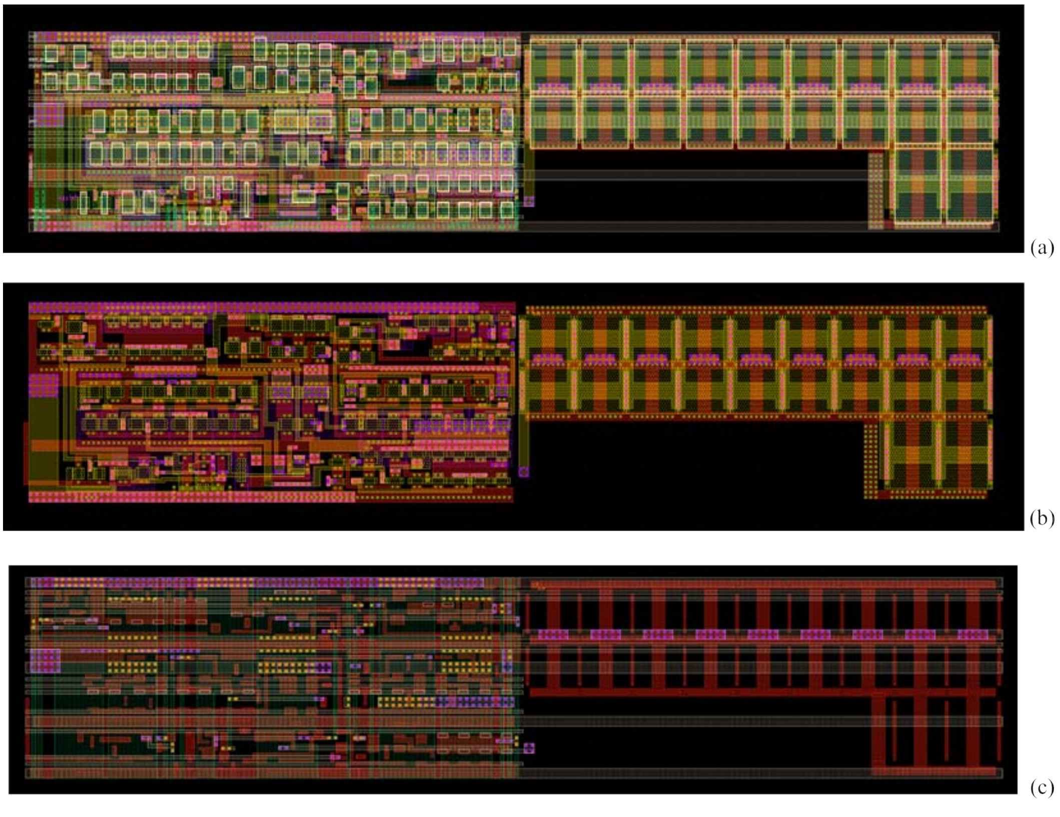

Figure A.1: The layout of schematic of Figure 4.5. 2.5pF NMOSCAP included but neuron and synapses excluded. **(a)** all layers. **(b)** Layers: N-well, Polysilicon, Metal 1, Metal 2, Vias. **(c)** Metal 1, Metal 3, Metal 4, Vias

BIBLIOGRAPHY

[1] C. A. Mead, *"Analog VLSI and Neural Systems,"* Addison-Wesley: Reading, MA, 1989.

[2] G. Indiveri, B. Linares-Barranco, T. Hamilton, A. van Schaik, R. Etienne-Cummings, et al., *"Neuromorphic silicon circuits,"* Frontiers in Neuroscience vol. 5, pp. 1–23, May 2011.

[3] E. Chicca, F. Stefanini and G. Indiveri, *"Neuromorphic electronic circuits for building autonomous congitive systems,"* proceeding of the IEEE, vol. X, no. x, XX. (Under review process)

[4] G. Turrigiano and S. Nelson, *"Homeostatic plasticity in the developing nervous system,"* Nature Reviews Neuroscience, vol. 5, pp. 97–107, February 2004.

[5] C. Bartolozzi and G. Indiveri, *"Global scaling of synaptic efficacy: Homeostasis in silcon synapses,"* Neurocomputing, vol. 72, no. 4-6, pp. 726-731, January 2009.

[6] M. O'Halloran and R. Sarpeshkar, *"A 10-nW 12-bit Accurate Analog Storage Cell With 10-aA Leakage,"* IEEE journal of solid-state circuits, vol. 39, no. 11, November 2004.

[7] M. O'Halloran and R. Sarpeshkar, *"An Analog Storage Cell with* $5e^{-}/sec$ *Leakage,"* IEEE International Symposium on Circuits and Systems, pp. 560-564, May 2006.

[8] K. Roy, S. Mukhopadhyay and H. Mahmoodi-meimand, *"Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuit,"* Proceeding of the IEEE, vol. 91, no. 2, pp. 305-327, February 2003.

[9] B. Linares-Barranco and T. Serrano-Gotarredona, *"On The Design and Characterization of Femtoampere Current-Mode Circuits,"* IEEE journal of solid-state circuits, vol. 38, no. 8, pp. 1353-1363, August 2003.

[10] C. Bartolozzi, S. Mitra, G. Indiveri, *"An ultra low power current-mode filter for neuromorphic systems and biomedical signal processing,"* IEEE Biomedical Circuits and Systems Conference, pp. 130-133, November 2006.

[11] C. Bartolozzi, O. Nikolayeva, G. Indiveri, *"Implementing homeostatic plasticity in VLSI networks of spiking neurons,"* IEEE International Conference on Electronics, Circuits and Systems, pp. 682-685, August-September 2008.

[12] S. C. Liu, B. A. Minch *"Homeostasis in a Silicon Integrate and Fire Neuron,"* Advanced Neural Information Processing Systems 2001, pp. 727-733.

[13] A. G. Andreou and K. A. Boahen, *"Translinear circuits in subthreshold MOS,"* Analog Integrated Circuits and Signal Processing, vol. 9, no. 2, pp. 141-166, March 1996.

[14] R. Wang, T. J. Hamilton, J. Tapson, A. van Schaik, *"An Analogue Memory for Spiking Neural Networks with 100-aA leakage,"* Transactions on Biomedical Circuits and Systems, vol. XX, pp. XX, XX. (Under review process)

[15] S. C. Liu, J. Kramer, G. Indiveri, T. Delbrück, and R. Douglas, *"Analog VLSI-Circuits and Principles,"* MIT Press, 2002.

[16] J. P. A. Pérez, S. C. Pueyo, B. C. López *"Automatic Gain Control,"* Springer New York, 2011.

[17] J. Mulder, W. A. Serdijn, A. C. van der Woerd, A. H. M. van Roermund *"Dynamic Translinear and Log-Domain Circuits: Analysis and Synthesis,"* The Springer International Series in Engineering and Computer Science, 1999.

[18] Y. Tsividis *"Operation and Modeling of the MOS Transistor,"* 2nd edition, Oxford University Press, 1999.

[19] D. Purves, *"Neuroscience,"* 4th edition, Sinauer Associates, 2007.

[20] E. Rodriguez-Villegas, A. J. Casson and P. Corbishley *"A Subhertz Nanopower Low-Pass Filter,"* IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 58, no. 6, pp. 351-355, 2011.

[21] F. Gozzini, G. Ferrari, and M. Sampietro, *“Linear transconductor with rail-to-rail input swing for very large time constant applications,”* Electronics Letters, vol. 42, no. 19, pp. 1069–1070, 2006.

[22] A. Becker-Gomez, U. Cilingiroglu, and J. Silva-Martinez, *“Compact sub-Hertz OTA-C filter design with interface-trap charge pump,”* IEEE Journal of Solid-State Circuits, vol. 38, no. 6, pp. 929–934, 2003.

[23] P. Bruschi, G. Barillaro, F. Pieri, and M. Piotto, *“Temperature stabilised tunable Gm-C filter for very low frequencies,”* Proceeding of the 30th European IEEE Solid-State Circuits Conference, pp. 107-110, 2004.

[24] R. Hogervorst, J. T. Tero, R. G. H. Eschauzier and J. H. Huijsing *"A Compact Power-Efficinet 3V CMOS Rail-to-Rail Input/Output Operational Amplifier for VLSI Cell Libraries,"* IEEE Journal of Solid-state circuits, vol. 29. no. 12, December 1994.