Università degli Studi di Torino
DIPARTIMENTO DI FISICA
Corso di Laurea Magistrale in Fisica

Tesi di Laurea Magistrale in Fisica

Front-End amplifiers in 65nm CMOS technology for the upgrade of the pixel detector of the CMS experiment

Candidato: Ennio Monteil

Relatore: Dott. Angelo Rivetti

Anno accademico 2012-13

Sommario

La tesi riguarda lo sviluppo di un chip innovativo per un rivelatore a pixel in tecnologia CMOS 65nm per l'upgrade dell'esperimento CMS previsto per l'High Luminosity Large Hadron Collider (HL-LHC) al CERN. La luminosità dell'acceleratore sarà aumentata di quasi un ordine di grandezza rispetto alla situazione attuale. Lo scopo principale dell'esperimento CMS è lo studio del bosone di Higgs e lo sarà anche all' HL-LHC, presso il quale misure di precisione saranno possibili. Per fare questo, è importante mantenere un rivelatore a pixel efficiente e performante anche nelle condizioni maggiormente ostili determinate da HL-LHC. Il rivelatore e la tecnologia attuale non sono sufficienti, pertanto un nuovo $R\&D$ è stato avviato sia sull'elettronica che sul sensore. La tesi descriverà in particolare il lavoro sulla catena analogica del nuovo chip per i pixel. I rivelatori a pixel, negli ultimi decenni, sono stati il motore dello sviluppo della tecnologia in HEP, dovendo integrare un'elettronica sofisticata in uno spazio piccolo, fronteggiando alti rate di dati per unità di area e un sensore altamente granulare. I futuri upgrade di LHC stanno spingendo ulteriormente i requisiti per il chip di lettura del pixel. Le principali sfide sono: pixel più piccoli per risolvere le tracce lasciate dalle particelle, più elevati rate di interazione ($1 - 2GHz/cm^2$), tolleranza alle radiazioni senza precedenti (10 MGy), una banda di trasmissione fino ad alcuni GB/s per chip. Inoltre, è importante sviluppare un'elettronica analogica e digitale a bassa dissipazione di potenza, con lo scopo di coprire grandi aree mantenendo basso il quantitativo di materiale. In particolare l'elettronica di Front-End deve avere delle performance a basso noise, essendo al contempo molto compatta e mantenendo un basso consumo di potenza per unità di area. L'argomento di questa tesi è la progettazione di un preamplificatore di front-end, che è il primo blocco della parte analogica. Il lavoro di tesi è organizzato come segue:

- Nel capitolo 1 viene data una panoramica sui rivelatori al silicio e sulle loro proprietà, descrivendo il rivelatore a pixel di CMS e i relativi upgrade.
- Il capitolo 2 riguarda il transistor MOS. La prima parte del capitolo è dedicata alle principali proprietà di questi dispositivi. La seconda parte, invece, riguarda le tecnologie deep submicron CMOS, largamente utilizzate nei circuiti

elettronici per rivelatori di radiazione. Attraverso un confronto tra le tecnologie 65 nm CMOS e 250 nm CMOS, utilizzata nell'attuale chip di lettura, vengono presentati i problemi tipici di queste tecnologie.

- Nel capitolo 3 sono descritti gli elementi di base di un front-end ideale e di i problemi relativi all'implementazione reale, dando particolare attenzione al rapporto segnale-rumore.

- La tecnica di ottimizzazione del noise è riportata nel capitolo 4. Essa viene inoltre applicata a un Charge Sensitive Amplifier, mostrando attraverso le simulazioni al CAD le scelte che permettono la minimizzazione del noise.

- Nel capitolo 5 vengono confrontati due Charge Sensitive Amplifier più realistici. Per ognuno sono riportate le simulazioni al CAD della potenza consumata, della linearità, del noise e dell'analisi dei mismatch. Viene inoltre studiata la compensazione della corrente di leakage del sensore.

Abstract

The thesis is focused on the development of an innovative chip for a pixel detector, using a CMOS 65nm technology, for the needed upgrade of the CMS experiment foreseen for the High Luminosity Large Hadron Collider (HL-LHC) at CERN. The luminosity of the accelerator will be increased of almost an order of magnitude with respect to the present.

The main purpose of the CMS experiment is the study of the Higgs boson and this will remain one of the main goals also at HL-LHC, where precision measurements will be possible. To do so, it is important to maintain an efficient and performant pixel detector even in the more hostile conditions determined by HL-LHC. Present detector and technology are not sufficient, so a new $R\&D$ has been launched on both the readout electronics and the sensor of the pixel and the thesis will describe in particular the work on the analog chain of the new pixel chip.

Pixel detectors, in the last decade, have been technology drivers in HEP, having to integrate sophisticated electronics in a small space, coping with high data rates per unit area and a very granular sensor. Future upgrades of LHC are pushing further the requests on a pixel chip. The challenges include: smaller pixels to resolve tracks in boosted jets, much higher hit rates ($1 - 2GHz/cm^2$), unprecedented radiation tolerance (10 MGy), much higher output bandwidth up to few GB/s. Furthermore, it is important to develop highly performant and low power analog and digital electronics, in order to instrument large areas while keeping the material budget low. In particular the very front-end electronics must have low noise performance, while being very compact and maintaining a low power consumption per unit area.

The subject of this thesis is the design of a Front-End preamplifier, the first block of the analog part. The thesis work is organized as follows:

- In chapter 1 an overview about silicon detectors and their properties is given. Furthermore, the CMS pixel detector is described, focusing also on the upgrades and the related requirements.

- Chapter 2 is about MOS transistors. The first part is focused on the main properties of these devices. The second part, instead, is about the deep submicron CMOS technologies, largely used in electronics circuits for radiation sensors. Through a comparison between the 65 nm CMOS and the 250 nm CMOS technology, used in the present readout chip, the issues typical of these technologies are presented.

- In chapter 3 the basic elements of a front-end chain are described in order to explain the key principles driving the design. Consequently, the issues which occur in real circuits are presented. Particular attention is given to the signal-to-noise ratio, which is a primary parameter of any front-end circuit.

- In chapter 4 the noise optimization technique in deep submicron CMOS technologies is reported. In addition, this methodology is applied to a first simple Charge Sensitive Amplifier architecture. The results of the CAD simulations show the best choice which minimizes the noise contribution.

- In chapter 5, finally, two more complete Charge Sensitive Amplifier architectures are compared. For each of them the CAD simulations of the power consumption, gain linearity, noise performance and the mismatch analysis are carried-out. Furthermore, the compensation of the detector leakage current is also studied.

Table of contents

Chapter 1

Particle detection in high energy physics

The first part of this chapter deals with the basic concepts of particle detection, focusing in particular on silicon detectors and their properties. The second part is about the CMS experiment, located at the LHC at CERN. A large section of this chapter is about the CMS pixel detector, giving particular attention to the phase 2 pixel upgrade. The front-end architectures studied in this work, in fact, are designed in order to meet the requirements given by this subsystem.

1.1 Tracking detectors

A detector is a device which describes particles properties through the detection of charged or neutral particles. All the detectors are based on the same principle: the particle which goes through the detector volume leaves some energy which is converted in an electrical signal analyzed by an electronic circuit.

In order to identify a particle it is necessary to combine the information which comes from different detectors. In particular tracking detectors are used to measure the particle momentum through a magnetic field and to reconstruct the primary vertex of the interaction. This device is also able to perform a measurement of the energy loss per unit path-length (dE/dX) which is an additional information for particle identification.

In the sensor design a key parameter which has to be taken into account is the mechanism with which charged particles lose energy into materials. This phenomenon happens due to scattering processes with the particles of the detector layers.
The main process is the inelastic Coulomb scattering with the atomic electrons.

In the region of $0.1 < \beta\gamma < 10^4$ this energy loss is described by the Bethe-Bloch formula:

$$-<\frac{dE}{dx}>= Kz^2\frac{Z}{A\beta^2}\left[\frac{1}{2}ln\left(\frac{2m_ec^2\beta^2\gamma^2T_{max}}{I^2}\right)-\beta^2-\frac{\delta}{2}-\frac{C(I,\beta\gamma)}{Z}\right] \qquad (1.1)$$

with:

K	$4\pi N_{av}r_e^2m_ec^2 = 0.30707\ MeV \times cm^2$ with r_e classical electron radius
z	charge of the incident particle in units of electron charge
Z	atomic number of absorption medium (14 for Si)
A	atomic mass of absorption medium (28 for Si)
m_ec^2	rest energy of the electron
β	velocity of the incident particle in units of c
γ	Lorentz factor $(1-\beta^2)^{\frac{1}{2}}$)
I	mean excitation energy (137 eV for Si)
T_{max}	maximum kinetic energy transfer in single collision

The term $(\delta/2)$ is a density effect correction and occurs at high energies due to a relativistic expansion of the electric field of the ionizing particle resulting in a slowdown of the logarithmic slope of the function. The term (C/Z) is a shell correction relevant at low energies because in this case electrons are not free and there is a screening effect for the most internal ones.

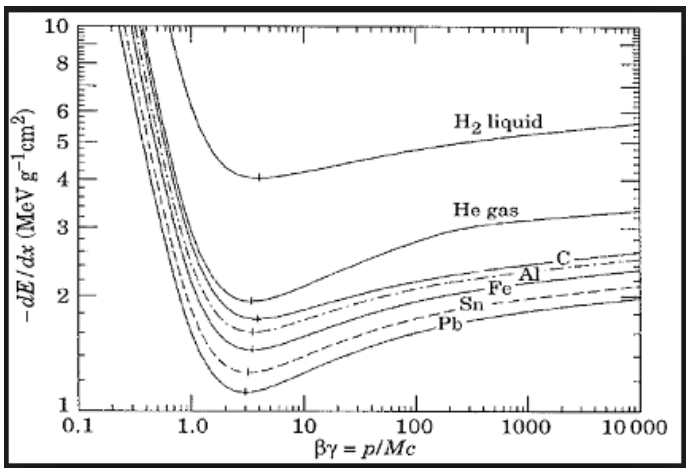


Figure 1.1: Energy loss for different particles

As one can see from Figure 1.1 the dE/dx is very high for low energy particles and decreases with the increase of particles energy until a minimum: a particle whose energy corresponds to this minimum value is called MIP (Minimum Ionizing

Particle). After the minimum the dE/dx increases with energy until it reaches a plateau for high energy particles.
In Figure 1.1 it is also possible to see that this equation is mass dependent especially in the low momentum region. This dependency can be expected for PID purposes.

1.1.1 Fundamental aspects

Every detector has some features which influence its behavior. In the following paragraphs a list of the main ones is presented.

- **Point resolution** The resolution of a detector is the minimum value of a physical quantity that it is possible to detect with the device. In the case of tracking detectors the target is to have a good vertex resolution: in high energy physics processes a lot of particles decay very little after the primary interaction, giving rise to secondary vertices which are very close to the interaction points.
  Another important aspect is the momentum resolution, which plays a key role in the particle identification. In order to measure the momentum, a magnetic field is applied in the detector, leading to a deflection of the charged particles, as described by the Lorentz force formula:

  $$F = q(\vec{v} \times \vec{B}) \tag{1.2}$$

  Equating the module of this force to the module of the centrifugal force:

  $$qvB = \frac{mv^2}{r} \tag{1.3}$$

  which, using $p = mv$ becomes:

  $$p = qBr \tag{1.4}$$

- **Occupancy** The occupancy of a detector refers to the number of busy channels, i.e. the number of channels hit by a particle. It is expressed as a percentage. If the occupancy value is too high the tracking becomes ambiguous. Since the pixel detectors at LHC constitute the seed of tracking, it is necessary to have a very low occupancy, below 1%.

- **Dead time and pile-up** Another very important concept, strictly related to the occupancy is the dead time(Δt), which is the time that the detector takes to process an event and mainly depends on the front-end electronics. It is crucial to reduce it as much as possible in order to increase the speed (and then the number of processed events) and because during the dead time the detection of the next event is not correct. It is possible to identify two different situations:

  - If the detector remains *unsensitive* during the dead time the subsequent event will be lost
  - If the detector remains *sensitive* the pile-up of the events occurs, leading to a distortion of the signal and loss of information about the two events.

- **Radiation hardness** High energy physics experiments are characterized by a large amount of radiation, which has to be tolerated especially by the tracking detectors, which are the most internal ones. It is a fundamental requirement which has to be fulfilled in order to have sensors which can work for a long time (some years) without losing too much in performance.

Regarding the tracking detectors, the best choice to meet these requirements is the use of semiconductor sensors, which are usually made of silicon.

1.2 Silicon sensors

Silicon (A=28, Z=14) is a semiconductor material with a bandgap of 1.12 eV, and the energy required to create an electron-ion pair is 3.6 eV. Another important property is the high mobility of the charged carriers which allows a very fast charge collection time.

As described in paragraph 1.1, particles energy loss is described by the Bethe-Bloch formula. This energy loss leads to the formation of a number of electron-ion pairs proportional to the amount of energy. It is possible to calculate the amount of pairs created by a MIP using the following formula

$$Q_{average} \simeq 100 \frac{q_e}{\mu m} d \tag{1.5}$$

with:

Q_e charge of an electron
d sensor thickness in μm

so for example for a thickness of 250 μm, which is one of the typical values used, around 25000 electron-hole pairs are produced.

It is important to underline that this is only an average value of a Landau distribution, which has a remarkable tail, so the most probable value is different from the average value. The fluctuation around the maximum is more important if the sensor is thinner: this situation has to be taken into account in the readout circuit design because it affects the dynamic range, i.e. the maximum allowable

voltage swing divided by the total noise voltage in the band of interest [1]. The main reason of the fluctuation is the production of the δ-electrons which is a quite rare but significant process, because their direction is perpendicular to the direction of the ionizing particle leading to irregular charge clouds and worse spatial resolution.

1.2.1 Signal formation

Silicon detectors are realized with different technologies but all of them are based on the same principle: they consist of a reverse bias pn junction. In a normal pn junction there is a transition between n-doped and p-doped material, in which some majority charges of one side diffuse into the differently doped side due to concentration difference. For this reason there is a recombination which leads to the formation of a region, close to the junction, where there are no free charge carriers. Therefore, it is called "depletion region". [2]
The sensor junction is reversely biased, in fact an electric field is applied to the two electrodes leading to the largest possible increase of the width of the depletion region. If a free charge enters this region, it will be removed by the electric field, so the charge carriers produced by a ionizing particle drift towards the electrode with the most favorable potential, producing a current in both the electrodes which can be read out.

1.2.2 Different types of silicon detectors

In particle physics different types of silicon detectors are used. The feature which mainly differentiate one of them by another is the shape of the electrodes. For example, silicon strip detectors are segmented in long and narrow elements, the strips, and each of them is an independent reverse-biased pn junction, while pixel detectors are segmented in smaller square elements. For the purposes of this work it is useful to investigate the pixel detector behavior in detail.

Silicon pixel detectors

Pixel detectors have had a large development in the last decades thanks to the high granularity due to the small dimensions of the single sensing element. The hybrid pixel detector, whose structure is shown in Figure 1.2, is a largely used type of silicon detector. The surface of the detector is segmented into small squares and the electronics has the same geometry. For each pixel an electronic chain provides amplification and some other functions such as data storage. Sensor and electronics are fabricated separately and subsequently connected through the "bump bonding"

technique, which consists of a metal deposition between the sensor output and the electronics input [3].

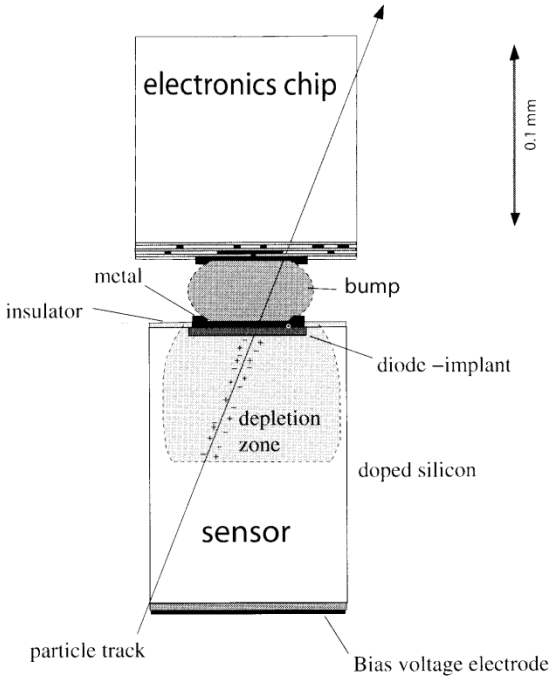
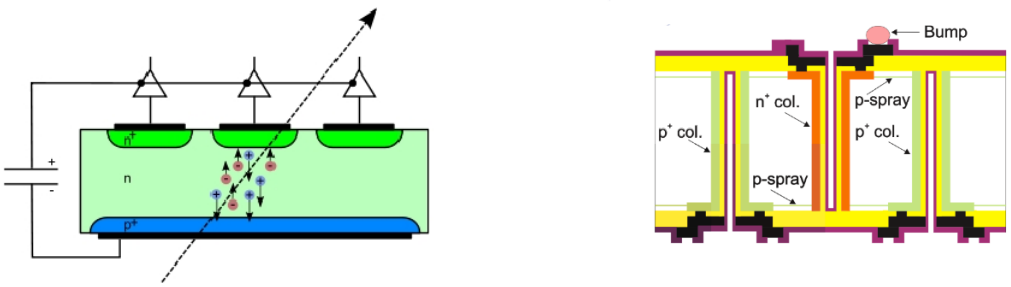


Figure 1.2: Hybrid pixel detector

Two notable hybrid pixel detectors tipologies are planar silicon sensors and 3D silicon sensors, shown in 1.3. The former ones are, right now, the most largely used, while the latter ones have experienced a large development during these last years.



(a) Planar (b) 3D

Figure 1.3: Planar silicon sensor vs 3D silicon sensor

The main difference between the two configurations is the electrodes arrangement. In the planar geometry, the electrodes are built on the top and bottom surfaces of the sensor. Otherwise, in the 3D geometry electrodes enter the sensor volume, perpendicularly to the surface [3].

1.3 The CMS experiment

CMS (Compact Muon Solenoid) is one of the experiment located at the LHC (Large Hadron Collider) at CERN. LHC is designed to reach collisions at a very high energy (around 7 TeV in the center of mass). The main purpose of the CMS experiment is the discovery and the study of the Higgs boson. Aside from it the experiment deals also with other particle physics topics, such as heavy ions physics. The CMS experiment is composed of some different detectors such as muon chambers(MUON), hadronic calorimeters(HCAL and HF), electromagnetic calorimeter (ECAL) and tracking system. This choice is made because, in order to achieve a good global event reconstruction, it is necessary to match the information which comes from the different detectors. The transverse section of the detector is reported below [4].

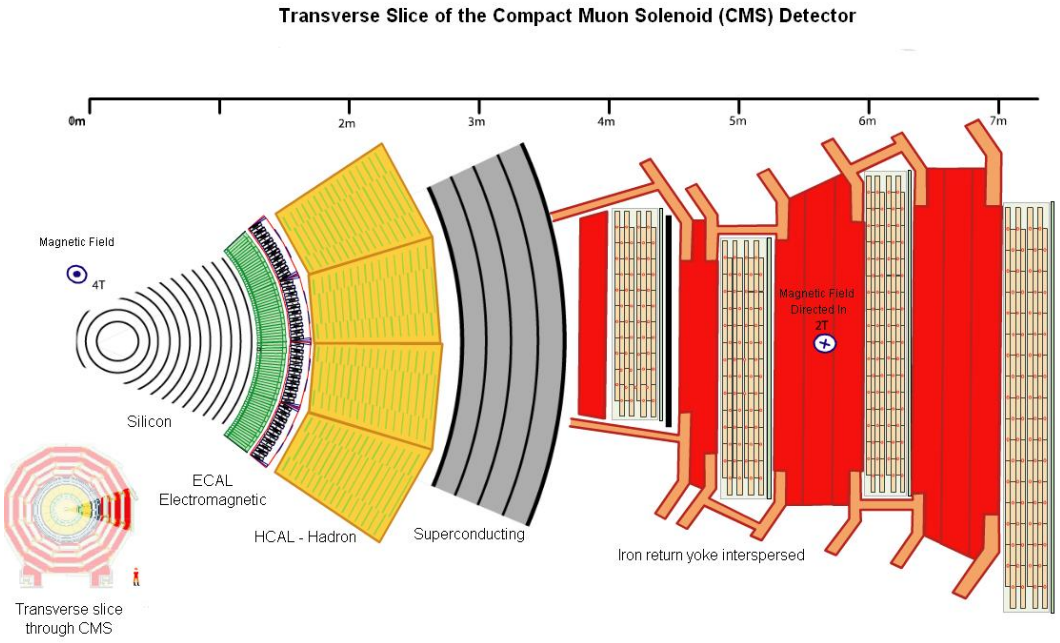


Figure 1.4: CMS detector

Considering the purposes of this work, in the next section the attention is focused on the tracking detector.

1.3.1 The CMS tracker

The CMS tracking system has been designed to provide precise measurements of the particle trajectories and reconstruction of the secondary vertices. The very special working conditions of the system played a key role in the choice of the detector layout. [5]

At the LHC design luminosity of $10^{34} cm^{-2} s^{-1}$, there are on average 1000 particles from more than 20 overlapping proton-proton interactions traversing the tracker per each bunch crossing(i.e. every 25ns). This situation leads to the requirement of a detector with high granularity and fast response in order to identify the trajectories and attribute them to the correct bunch crossing. Another issue is the severe radiation damage that the detector experiences due to the intense particle flux. The main challenge in the design of the tracking system was to develop detector components able to operate in this harsh environment for a long lifetime.
It is not so simple to achieve all these goals: these features also imply a high power density of the on-detector electronics, and therefore an efficient cooling system: the detector operating temperature, in fact, has to be below -10°C in order to minimize radiation damage. Nevertheless, this system has to be implemented with the smallest possible amount of material, in order to minimize multiple scattering, bremsstrahlung, photon conversion and nuclear interaction. In the detector design it is thus necessary to find a compromise with this respect.

These requirements on granularity, speed and radiation hardness lead to a tracker design entirely based on a silicon detector technology. The current CMS tracker is composed of a pixel detector with three barrel layers and a silicon strip tracker with 10 barrel detection layers. Each system is completed by endcaps which consist of 2 disks in the pixel detector and 3 plus 9 disks in the strip tracker on each side of the barrel, extending the acceptance of the tracker up to a pseudorapidity of $|\eta| < 2.5$. The system surrounds the interaction point with a diameter of 2.5m.

1.3.2 Pixel detector

The pixel system is the part of the tracking system that is closest to the interaction region. It contributes precise tracking points in $r - \phi$ and z and therefore is responsible for a small impact parameter resolution, essential for good secondary

vertex reconstruction.

Current pixel detector

The detector layout is shown in Figure 1.6. The sensors for the CMS pixel detector adopt the n-on-n concept. The pixels consist of high dose n-implants introduced into a high resistance n-substrate. Despite the higher costs due to the double sided processing this concept was chosen as the collection of electrons ensures a high signal charge at moderate bias voltages ($< 600V$) after high hadron fluences. Furthermore the double sided processing allows a guard ring scheme keeping all sensor edges at ground potential.
It is interesting to observe that, during data taking, more than 95% of the pixel channels are active [6] and, thanks to its high segmentation, this detector is used not only in the offline track reconstruction but also in the online tracking performed by the HLT (High Level Trigger). [7]
This detector was designed for a maximum luminosity of $1 \times 10^{34} cm^{-2} s^{-1}$, at which the electronics chip, the PSI46v2, has a dynamic inefficiency around 4% so it suffers an important data loss as the luminosity increases over this value.
Focusing on radiation hardness, at the specified luminosity the innermost pixel layer, which has a radius of 4.4 cm, experiences a particle fluence of $3 \times 10^{14} n_{eq}/cm^2/yr$, but all the pixels are designed to tolerate even a double radiation with respect to this one. However, with the increase of luminosity, also radiation hardness will become an issue.

LHC development

In 2013 the first Long Shutdown (LS1) of LHC has begun. Its aim is to improve the machine performance, especially in terms of luminosity and beam energy. Now, the collider luminosity is $7 \times 10^{33} cm^{-2} s^{-1}$ and the beam energy is around 3.5 TeV. In 2015, when the collider will resume operation, it will get the design luminosity ($10^{34} cm^{-2} s^{-1}$) and the beam energy will increase up to around 6.5 TeV. This fact also requires an upgrade of the detector, because it will no be longer able to tolerate these conditions. To change this situation a new version of the pixel detector (phase 1 upgrade) will be introduced between 2016 and 2017.

LHC will be further upgraded and in 2022-2023 the HL-LHC (High Luminosity LHC) will be built. Its name comes from the fact that the luminosity of the machine will be increased up to $5 \times 10^{34} cm^{-2} s^{-1}$ at least, while the beam energy will be around 7 TeV. The increased environment hostility will require a further detector upgrade (phase 2 upgrade).

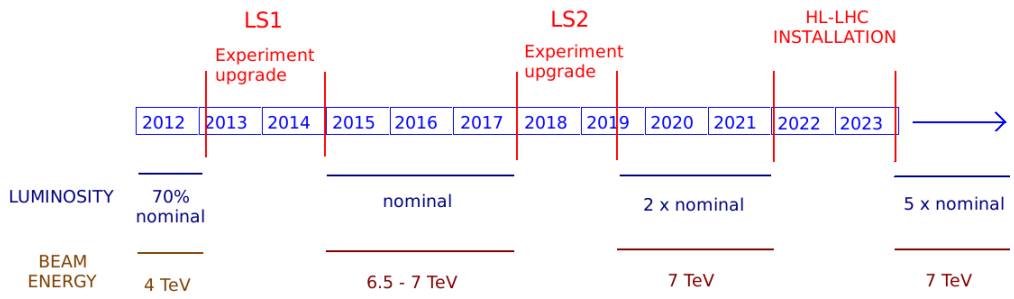


Figure 1.5: LHC timeline

Pixel detector upgrade phase 1

In order to have a detector which maintains or even improves its performance despite the issues which will play a key role in the next LHC runs, it becomes necessary to improve the detector design. In Figure 1.6 a comparison between the current pixel detector and the phase 1 detector is reported.

A first aspect which has to be taken into account is that the first detector layer has to be as close as possible to the beam. This requirement is achieved through the new LHC beam pipe, with a smaller radius. The new element will be introduced paying attention, on the other hand, not to worsen the operating conditions of the accelerator. For this reason a fourth pixel layer will be inserted in the detector.

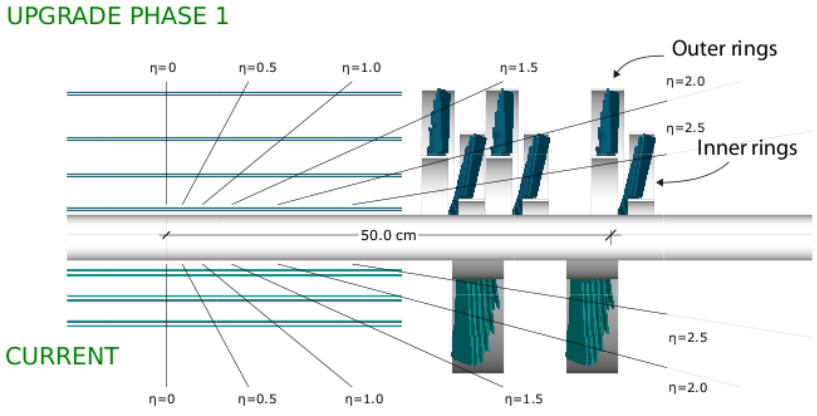


Figure 1.6: Current pixel detector vs phase 1 pixel detector

Furthermore, comparing the simulation studies of the proposed Phase 1 upgrade

pixel detector operating at a luminosity of $2 \times 10^{34} cm^{-2} s^{-1}$ with the present pixel detector, it is possible to see the key limitations of the current pixel detector that should be addressed by the upgrade:

- *Data loss at high occupancy and trigger rate* The purpose is to use a pixel readout chip (ROC) able to tolerate luminosities up to 2×10^{34} $cm^{-2} s^{-1}$ and pile up to 50 [7] maintaining a high performance and a low fake rate. The current detector, instead, suffers from a significant data loss in these conditions, also due an insufficient readout speed.

- *Material budget and radiation damage* In the phase 1 upgrade a new design will be adopted with a different cooling system and electronics boards and connections located out of the interaction volume. The amount of material is, in fact, a major issue due to multiple scattering but not only: some particles can be lost in nuclear interactions. Also the radiation damage is expected to be diminished. In fact, although the nominal radiation damage will be approximately the same, the lower charge threshold for pixel hits mitigates the effects of the reduced collection charge.

Pixel detector upgrade phase 2

This upgrade will be performed around 2022-2023, when the LS3 will occur, giving rise to the High Luminosity (HL) LHC. This LHC upgrade results in higher machine performances with which the pixel system has to cope: a luminosity of $5 \times 10^{34} cm^{-2} s^{-1}$, a pile-up of 140 (in 25 ns) and an integrated luminosity of $270 fb^{-1}$ per year [6].
The goals of the upgrade are:

- Higher radiation hardness of inner layers
- Increased granularity using smaller pixels
- Improved rate capability of the ROCs
- Contribution to level 1 trigger
- Lower power consumption

From the geometry point of view the choice will be very similar to the phase 1 geometry, with 4 pixel layers in the barrel, but with more disks in the outer part in order to improve the physics at high η values. The pixel size will be reduced in order to have an improvement in granularity and, consequently, in spatial resolution. The sensor choice has not yet been made. For the outer layers planar silicon sensors

will be definitely used, but for the inner ones there is the chance of using 3D silicon sensors, if their performances will be considered better than the planar ones.
Also the ROC optimization depends of several factors: the pixel size, hit rates, data buffering and readout, but also radiation hardness. It has in fact to tolerate around 300 MRad in 10 years. From the preliminary studies it has been understood that the new chip has to sustain particle rates up to $2GHz/cm^2$, taking into account the huge statistical fluctuations expected.

PARAMETER or FEATURE	**1st generation LHC phase 0**	**2nd generation LHC Phase 1**	**3rd generation LHC Phase 2**
Max Particle Flux	~50 MHz/cm^2	~200 MHz/cm^2	~500 MHz/cm^2
Max Pixel Flux	200 MHz/cm^2	600 MHz/cm^2	2 GHz/cm^2
Rad. Hardness	1.5 MGy	3.5 MGy	10 MGy
Pixel Dimension	100x150 mm^2 50x400 mm^2	100x150 mm^2 50x250 mm^2	25x150 mm^2 50x100 mm^2
Signal Threshold	2500-3000 e$^-$	1500-2000 e$^-$	~1000 e$^-$
L1 Trigger Latency	2-3 us	4-6 us	6-20 us
L1 Trigger Rates	100 KHz	~100 KHz	200-1000 kHz
L1 Trigger contribution	no	no	clustering info @L0 self-triggering
ASIC side	~1 cm^2	~4cm^2	1-4cm^2
Hit memory per chip	0.1 Mb	1 Mb	~16 Mb
Chip output bandwidth	~40 Mb/s	~320 Mb/s	~3 Gb/s
Power Budget	~0.3 W/cm^2	~0.3 W/cm^2	<0.4 W/cm^2
Electronics technology node	250nm CMOS	250nm CMOS (CMS) 130nm CMOS (ATLAS)	65nm CMOS

Figure 1.7: Detector specifications

Given these requirements a different choice has been made about the electronics technology: the *65nm CMOS* will replace the *250nm CMOS*. It will be done due to the advantages given by the 65nm in terms of:

- **High density** It gives the chance of building complex circuits is a small area, thanks to the smaller transistor dimensions available
- **Radiation tolerance** It has been already verified that 65nm technology transistors are able to tolerate up to 200 MRad of radiation, and there are ongoing tests also for higher values
- **Speed and low power**
- **Long term availability** It is a mature technology and thanks to the fact that it is used in many applications, it will have a long-term availability, which is

an important achievement given that the HL-LHC will operate for around 10 years starting in 2023.

Chapter 2

Deep submicron MOS

In this chapter a general review about the main features of the MOS transistor, the fundamental building block of all the circuits presented in this work, is given. After an introduction about silicon, the material the transistor is made of, the first part of the chapter is focused on the MOS transistor and its behavior, with a particular attention on the electrical characteristics. The second part, instead, is dedicated to the deep submicron CMOS technologies, with a focus on the fact that in these cases it is necessary to cope with new issues, the "short channel effects". Lastly, since the phase 2 pixel upgrade will use a *65nm* CMOS technology in place of the *250nm*, a comparison between these two technologies has been performed, in order to point out advantages and disadvantages of the different choice.

2.1 NMOS and PMOS

Before starting with the description of the MOS transistor it is appropriate to enumerate the silicon properties which make advantageous the use of this material in the transistor fabrication.

2.1.1 Silicon properties

Silicon is a tetravalent element, i.e. it has four electrons in the more external level. It is a semiconductor with a band gap energy of 1.12 eV at $T = 300K$.

Silicon is not the only semiconductor that can be chosen to fabricate transistors. Nevertheless it is the most used thanks to its particular features:

- The abundance of this element
- The easiness of growing high-purity silicon crystals

- The possibility of taking advantage of the electrical properties of silicon through the doping technique, i.e. introducing in the crystal lattice some atoms of other chemical elements
- Its thermical and mechanical properties: silicon devices can operate in a large interval of temperature, usually identified between -55°C and 125°C
- It allows the use of photo-lithographic techniques

2.1.2 NMOS transistor

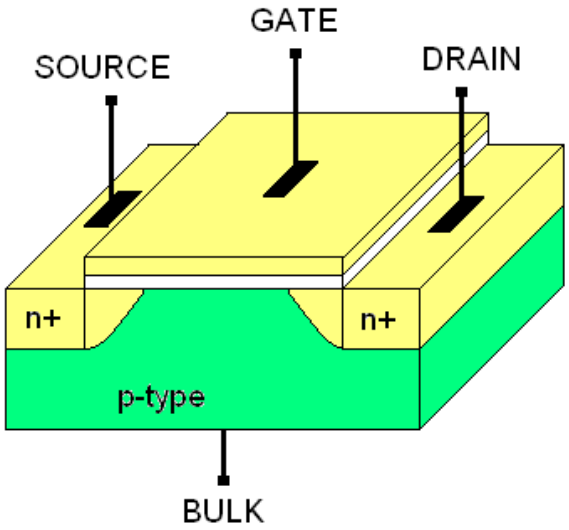


Figure 2.1: NMOS transistor

Figure 2.1 outlines the structure of a NMOS transistor. It consists of a $p-doped$ region, which constitutes the bulk or substrate, on which two $n^+-doped$ electrodes are implanted. It is important to underline that the distinction between the source and drain terminals indicated in the image is due to the potential applied to the electrodes, which otherwise are identical. As one can see, the electrodes and the substrate form two pn junctions which have to be reverse biased so that the transistor works correctly. The part between source and drain is instead called "channel region". Just over it there is a thin (typical values are in the range 2-7 nm, depending on the technology) layer of oxide (SiO_2) to separate the channel from the gate, which is the control electrode. The gate is in general made of polysilicon. However, due to the fact that it is grown on a oxide, it does not form a uniform crystal, but it is made of grains with different crystal orientation, hence the name "polysilicon"

or, in short, "poly".

Let's go now to analyze the main parameters which must be taken into account during the device fabrication. Besides the oxide thickness, fundamental quantities are the channel length (L), which is the distance between source and drain, and the channel width (W), the dimension orthogonal to the L. In reality, however, it happens that the two electrodes widen a bit in the channel region, each one by a value L_D, resulting in a decrease of the channel length: a relevant quantity is then the effective channel length given by:

$$L_{eff} = L - 2L_D \tag{2.1}$$

Nevertheless, bearing in mind these considerations, usually in literature L_{eff} is simply called channel length and is indicated with L. This parameter is however very important because usually it is the minimum channel length available which defines the technology used. As an example, in the case of sensors Front-End, the currently used technologies are between $0.35\mu m$ and $65nm$.

2.1.3 PMOS transistor

In Figure 2.2, instead, it is possible to see the structure of a PMOS transistor.

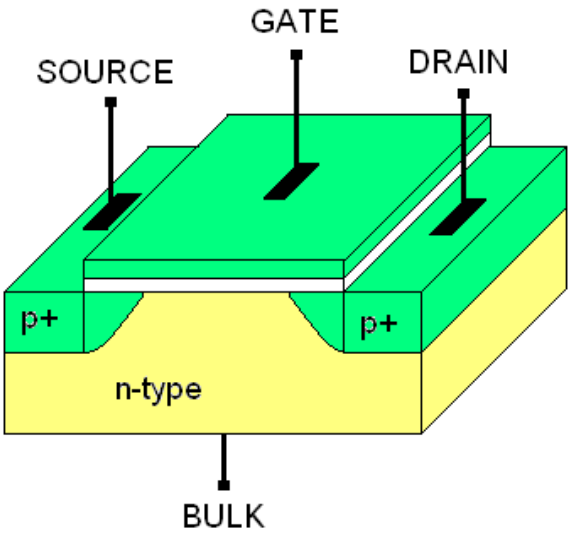


Figure 2.2: PMOS transistor

At this point the explanation of the PMOS operation is quite simple, because it is completely mirrored with respect to the NMOS. It explaines why it is common to call these process CMOS, i.e. Complementary MOS. In fact, this device consists of a n-doped substrate on which p-doped electrodes are built. In order to have positive quantities in the equation which describe its behavior, source and drain are reversed with respect to the NMOS.

2.1.4 MOS symbols

In the schematics reported in literature it is possible to find a lot of different representations of NMOS and PMOS transistors. In this paragraph only the most used are shown. For example they can be implemented with only three terminals: a very common depiction is the one reported below, in which NMOS and PMOS are distinguished thanks to the different arrow direction on their source terminal:

(a) NMOS (b) PMOS

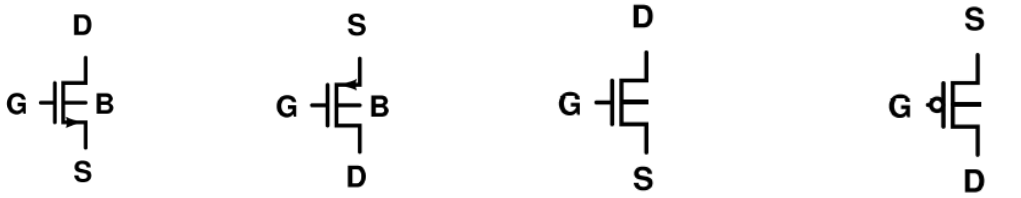
Figure 2.3: NMOS and PMOS with three terminals

Another frequently used representation is reported in 2.4. In this case the PMOS can be distinguished from the NMOS by the presence of a small circle next to the gate terminal:

(a) NMOS (b) PMOS

Figure 2.4: NMOS and PMOS with three terminals (alternative representation)

Of course both these representations can also take into account the presence of the bulk terminal, as shown in the following examples, resulting in a four-terminals symbol:



(a) NMOS (b) PMOS (c) NMOS (d) PMOS

Figure 2.5: NMOS and PMOS with four terminals

2.1.5 Success of CMOS technologies

The main advantage of the CMOS processes is the possibility of building both the NMOS and the PMOS on the same substrate. It is usually preferred the usage of a p-type substrate. Firstly, it can be produced more easily than a n-type one. Furthermore, due to a lower mobility resulting in a higher resistivity, it allows the limitation of noise propagation through the substrate in complex IC (Integrated Circuits). At this point it is necessary to understand how to produce a PMOS transistor, assumed that the substrate is p-doped whilst for a PMOS a n-doped one is needed. The idea is to realize a counter-doping in selected areas of the wafer, called nwells. Nevertheless, even if the wafer substrate is p-doped, also in the NMOS fabrication a pwell is realized in order to have a more precise doping profile. In addition, in order to guarantee a full electrical insulation betwen one device and another, STIs (Shallow Trenched Insulations) are used. This technique results in the removal of the silicon in which devices are not foreseen, replacing it with oxide. Lastly, it is useful to underline that the interconnections between the different devices are realized with metal lines. Figure 2.6 shows a typical structure realized following the previous description.

2.2 Electrical characteristics of CMOS transistors

It is now crucial to understand how NMOS and PMOS behave into an electronic circuit, considering their different operating regions. They basically depend on the bias voltages, i.e. the voltages applied to the device terminals. Keeping in mind that the PMOS behavior is mirrored with respect to the NMOS, the attention is

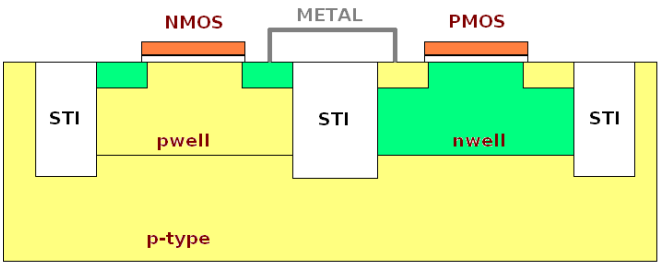


Figure 2.6: View of a CMOS process

focused on the latter.

2.2.1 Threshold voltage

The first situation which is interesting to analyze is when source, drain and bulk are at zero and only the gate voltage is changed. It is possible to identify some different situations:

- **Accumulation**: it occurs when the gate voltage is negative and, accordingly, holes are attracted underneath the gate oxide.
- **Depletion**: in this case the gate voltage is just above zero. Holes are now removed from the channel, in which a layer of ionized acceptor atoms is formed, but they can not move.
- **Inversion**: the gate voltage is further increased, giving rise to the presence of electrons free to move in the channel between source and drain.

It is common to choose the source terminal as the voltage reference, so the gate voltage is indicated as V_{GS} and the drain voltage as V_{DS}. In first approximation, in classical models, it is possible to say that free electrons are present in the channel only when the condition $V_{GS} - V_{TH} > 0$ is verified (strong inversion); nevertheless, especially in submicron CMOS technologies, the subthreshold region plays a key role in a lot of circuit implementations.

2.2.2 I/V characteristics

Considering the inversion region, also the V_{DS} voltage has to be taken into account, giving rise, in turn, to two different regions, as shown in the I_{DS} vs V_{DS} curve in Figure 2.7

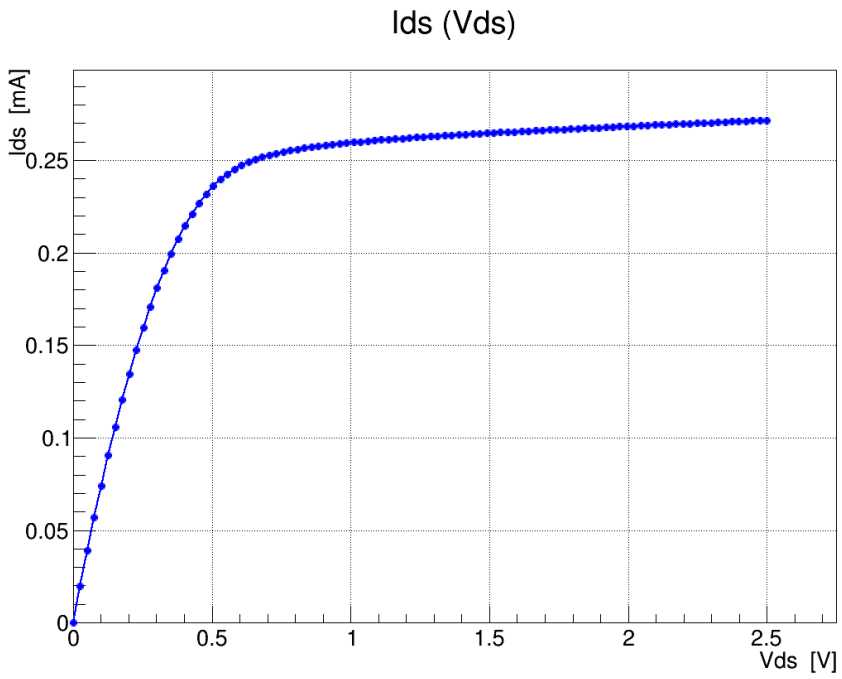


Figure 2.7: I_{DS} vs V_{DS}

- **Linear region** If V_{DS} is very small (up to around two hundreds of mV) the electric field through the channel is constant, leading to a linear relationship between I_{DS} and V_{DS}:

$$I_{DS} = \mu_n C_{ox} \frac{W}{L}(V_{GS} - V_{TH})V_{DS} \tag{2.2}$$

  But if V_{DS} is raised only a little more the charge density in the channel can not be considered uniform, leading to a modified expression with respect to the 2.2 :

$$I_{DS} = \mu_n C_{ox} \frac{W}{L}\left[(V_{GS} - V_{TH})V_{DS} - \frac{V_{DS}^2}{2}\right] \tag{2.3}$$

- **Saturation region** The Equation 2.3 shows that at a certain point a maximum is reached and, subsequently, according to the Equation 2.3 the current starts to fall. This is not a physical situation. In reality, in the points in which $V_{DS} > V_{GS} - V_{TH}$ the channel is not anymore inverted: it is usual to say that the channel is "pinched-off", and it does not contribute to the current flow with its own carriers. The current expression then changes in the following way:

$$I_{DS} = \frac{1}{2}\mu_n C_{ox} \frac{W}{L}(V_{GS} - V_{TH})^2 \tag{2.4}$$

The element which has to be underlined is that the current is independent from the V_{DS} voltage, so in the saturation region the transistor behaves as a voltage controlled current source.

However, it is more rigorous to underline that also in the saturation region a little dependence from V_{DS} remains due to the channel length modulation effect. The latter happens because, increasing V_{DS}, the pinch-off point moves towards the source, resulting in a decrease of the effective channel length. The current formula has therefore to be written as follows:

$$I_{DS} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2(1 + \lambda V_{DS}) \tag{2.5}$$

λ is the channel length modulation parameter.

2.2.3 Weak inversion operation

As technology has advanced, the behavior of the simple MOSFET has become more complicated, due to smaller geometries, higher electric fields and a continuing decrease in the power supply voltage. In this context the weak inversion, i.e. the current conduction in the subthreshold region, ($V_{GS} - V_{TH} < 0$), has gained more and more importance. It is interesting to observe that, in weak inversion operation, the I_{DS} dependence from V_{GS} is an exponential one, very different from the strong inversion case:

$$I_{DS} = 2n\mu C_{ox}\frac{W}{L}\phi_T^2 e^{\frac{V_{GS}-V_{TH}}{n\phi_T}}\left(1 - e^{-\frac{V_{DS}}{\phi_T}}\right) \tag{2.6}$$

Inversion coefficient (IC)

Having to discriminate between different regions of inversion, a more useful approach to the interpretation of the MOSFET behavior for analog design, based on the level of inversion (or Inversion Coefficient, IC), can be used. This method provides a convenient way of identifying the operating region and inversion level of MOS transistors, which can be used as a design variable for circuit optimization [8]. The definition of IC is the following:

$$IC = \frac{I_D}{2n\mu C_{OX}\left(\frac{W}{L}\right)\phi_T^2} \tag{2.7}$$

with:

I_D	drain current
n	slope factor
C_{OX}	gate oxide capacitance
ϕ_T	thermal voltage (=kT/q)
W	effective channel width
L	effective channel length

The different transistor operating modes are identified by the IC values:

- $IC > 10 \rightarrow$ **Strong inversion** i.e. if $V_{GS} > V_{TH}$
- $0.1 < IC < 10 \rightarrow$ **Moderate inversion** i.e. if $V_{GS} \simeq V_{TH}$
- $IC < 0.1 \rightarrow$ **Weak inversion** i.e. if $V_{GS} << V_{TH}$

Moreover, also another quantity, called "inversion factor" and indicated with the symbol γ, is relevant. The mathematical expression is:

$$\gamma = \frac{1}{2} + \frac{1}{6}\frac{I_C}{I_C + 1} \tag{2.8}$$

It will be used in the following chapters in the noise calculations.

2.2.4 Small signal parameters

The previous found characteristics are not linear, but in the circuits analysis it is very important to deal with linear quantities. Using Taylor series only at the first order approximation it is possible to write, for a function f, the following expression:

$$f(x) = f(x_0) + \left(\frac{\partial f}{\partial x}\right)_{x=x_0} (x - x_0) \tag{2.9}$$

If this function has a variation around x_0:

$$\Delta f = \left(\frac{\partial f}{\partial x}\right)_{x=x0} \Delta x \tag{2.10}$$

The importance of this formula is that it provides a linear relationship between variations that happen at an equilibrium point. Let's now list the main small signal parameters used in the circuit analysis.

Gate transconductance

The gate transconductance is the derivative of the I_{DS} current with respect to the V_{GS} voltage. The name comes from the fact that these two quantities are measured in different terminals. Using the strong inversion expression of the current:

$$g_m = \frac{\partial I_{DS}}{\partial V_{GS}} = \mu_n C_{ox} \frac{W}{L}(V_{GS} - V_{TH}) \tag{2.11}$$

From the previous relationship it is hence clear that g_m associates a variation of the I_{DS} current to a variation of the V_{GS} voltage.

Output conductance

Since the V_{DS} voltage plays a role in the current value due to the channel length modulation, it is useful to define a quantity which links a variation of the current to a variation of V_{DS}:

$$g_{ds} = \frac{\partial I_{DS}}{\partial V_{DS}} = \lambda \frac{1}{2} \mu_n C_{ox} \frac{W}{L}(V_{GS} - V_{TH})^2 \simeq \lambda I_{DS} \tag{2.12}$$

The last equality is approximated because in the I_{DS} also the channel length modulation would have been taken into account. It is important to underline that the reciprocal of g_{ds} is the output resistance, r_0:

$$r_0 \simeq \frac{1}{\lambda I_{DS}} \tag{2.13}$$

Lastly, it is necessary to stress that the output conductance plays a key role in the definition of the gain.

2.2.5 The g_m/I_D method

The g_m/I_D method is largely used in Integrated Circuit design. It considers the relationship between the ratio of the transconductance g_m over DC drain current I_D and the inversion coefficient as a fundamental design tool. Three main reasons lead to the choice of g_m/I_D: [9]

- It is strongly related to the performances of analog circuits
- It gives an indication of the device operating region
- It provides a tool for calculating the transistor dimensions

Considering a constant current value, the greater is the ratio g_m/I_D the greater is the transconductance value, so in other words it is possible to say that this ratio is a measure of the efficiency of translating current into transconductance.

Another important property of the g_m/I_D is that it is related to the transistor operating mode as we can see from the following formula:

$$\frac{g_m}{I_D} = \frac{1}{I_D}\frac{\partial I_D}{\partial V_G} = \frac{\partial(lnI_D)}{\partial V_G} = \frac{\partial\left\{ln\left[\frac{I_D}{\left(\frac{W}{L}\right)}\right]\right\}}{\partial V_G} \tag{2.14}$$

The derivative is maximum in the weak inversion region where the I_D dependence versus V_G is exponential while it is quadratic in strong inversion, becoming almost linear deeply in strong inversion. In Figure 2.8 it is shown that the g_m/I_D ratio decreases as the operating point moves toward strong inversion when I_D or V_G are increased.

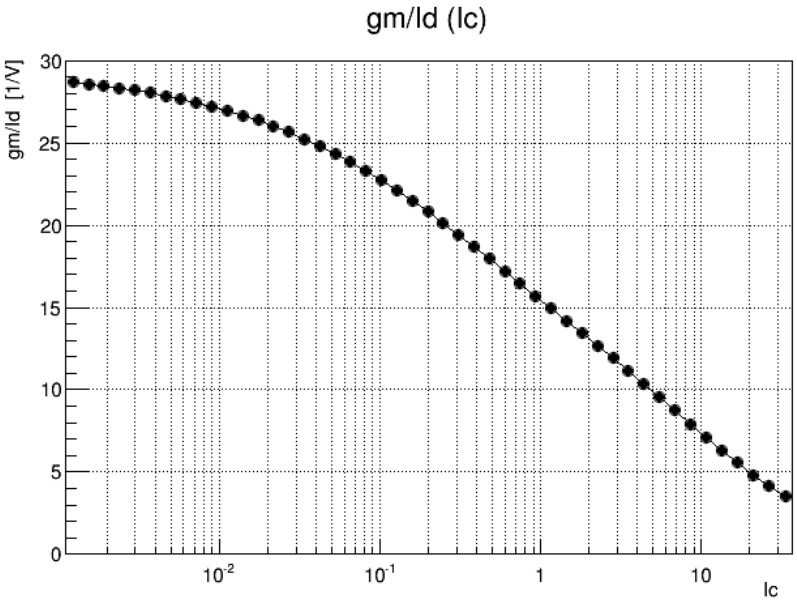


Figure 2.8: g_m/I_d for a NMOS transistor

2.3 Deep submicron CMOS technologies

In the last decades the number of transistors in electronics devices is tremendously grown, thanks to the transistor scaling which leads to the production of

smaller and smaller transistors. This element, together with the zero static power dissipation of CMOS logic, underlies the dominance of CMOS technologies [10]. But the scaling process is not so simple, because a lot of effects emerge when the channel length is approximately below $3\mu m$, affecting the transistor behavior: they are called "short channel effects". First of all, however, it is necessary to explain the scaling theory, i.e. how the transistor modifies its behavior with scaling.

2.3.1 Scaling theory

The ideal transistor scaling is marked by three principles [10]:

- Reduction of all the lateral and vertical dimensions by a factor α
- Reduction of the threshold and supply voltages by α
- Increase of all the doping levels by α

It is then important to underline that, due to the contemporary scaling of dimensions and voltages, all the electric fields in the transistor remain approximately constant. The scaling of these quantities obviously results in a change of the I_{DS} current equations. In strong inversion:

$$I_{DS,scaled} = \frac{1}{2}\mu\alpha C_{ox}\frac{W}{L}\left(\frac{V_{GS}}{\alpha} - \frac{V_{TH}}{\alpha}\right)^2 = \frac{I_{DS}}{\alpha} \tag{2.15}$$

Other interesting aspects are the behaviors of the transconductance and the channel capacitance, always in strong inversion:

$$g_{m,scaled} = \alpha\mu C_{ox}\frac{W}{L}\frac{(V_{GS} - V_{TH})}{\alpha} = g_m \tag{2.16}$$

$$C_{ch,scaled} = \frac{W}{\alpha}\frac{L}{\alpha}(\alpha C_{ox}) = \frac{1}{\alpha}WLC_{ox} \tag{2.17}$$

The transconductance does not scale because the decrease of the overdrive voltage is compensated by the increase in the density of the gate capacitance. The total channel capacitance, instead, decreases with scaling: this is a remarkable advantage, because it leads to an increase in speed and in a reduction of power dissipation.

2.3.2 Short-channel effects

The short-channel effects mainly happen due to [1]:

- Increase of the electric fields caused by a non-proportional power supply voltage scaling.

- The built-in potential is neither scalable nor negligible.
- Complexity in reducing the depth of Source-Drain junction.
- Decrease of the mobility due to the increase of the substrate doping.

It is then important to make a list of the main short-channel effects, since they have a great influence on the simulation of circuits developed with deep submicron CMOS technologies.

Threshold voltage variation

In scaled transistors the V_{TH} value is strongly affected by the channel length. In fact, the depletion regions associated to source and drain tend to extend a bit into the channel area, so the immobile charge imaged by charge on the gate is reduced, resulting in a lowering of the threshold voltage. It means that the threshold voltage increases with L. If L is very little, however, the widening of the depletion regions is so relevant that they can come in proximity with each other, filling almost all the channel. To fix this situation a non uniform doping implant, called "Halo implant", is performed near the depletion region. The halo implant, however, gives rise to the "Reverse Short Channel Effect (RSCE)": the threshold voltage is high when L is small, then increasing L the threshold voltage diminishes.

Mobility reduction

If the V_{GS} voltage is large, the constant-field configuration described before can not be applied. In fact, the high electric field between gate and channel results in a constriction of the region under the oxide in which carriers can move, leading to an increment of multiple scattering which reduces the mobility. This effect can be represented using an empirical equation:

$$\mu_{eff} = \frac{\mu_0}{1 + \theta(V_{GS} - V_{TH})} \tag{2.18}$$

μ_0 denotes the mobility value when the degradation is not present. Since the mobility expression is changed, also the current expression is modified:

$$I_{DS} = \frac{1}{2} \frac{\mu_0 C_{ox}}{1 + \theta(V_{GS} - V_{TH})} \frac{W}{L} (V_{GS} - V_{TH})^2 \tag{2.19}$$

Velocity saturation

It is necessary to underline that the mobility has also a dependence on the lateral electric fields in the channel, dropping for fields higher than $1V/\mu m$. The carrier

velocity is

$$v = \mu E \tag{2.20}$$

so it tends to saturates at around $10^7 cm/s$, leading to a I_{DS} saturation before the pinch-off point. The expression of the current is modified:

$$I_{DS} = WC_{ox}v_{sat}\frac{(V_{GS} - V_{TH})^2}{V_{GS} - V_{TH} + 2\frac{v_{sat}L}{\mu_{eff}}} \tag{2.21}$$

It is however useful to stress that if V_{sat} or L are large, the Equation 2.21 reduces to the well known square-law relationship, pointing out that this effect is relevant only with small channel lenghts.

Gate leakage current

In deep submicron technologies also the oxide thickness is scaled-down becoming very thin. Some electrons are hence allowed to tunnel through the oxide, giving rise to a small current. In some cases, especially when the oxide thickness is below 1.5 nm (luckily it happens for 45nm or below CMOS technologies) , this current can significantly rise, compromising the insulation between the gate and the channel. Also in 65nm technologies it is however important to carefully control the leakage current values, especially if many transistors are put in parallel.

Hot carrier effects

High V_{DS} voltage values result in large lateral electric fields. As a consequence, despite the saturation of the charge average velocity, the instantaneous velocity increases, and then also the kinetic energy, giving rise to the so called "Hot carriers". Near the drain these carriers can hit silicon atoms at high speeds producing impact ionization. In other words, it gives rise to an additional number of electrons and holes and, hence, to a drain-substrate current. In fact, electrons are absorbed by the drain and holes by the substrate. This issue has to be taken into account as it affects the output conductance behavior.

Output impedance variation with V_{DS}

Until now the assumption of a constant output impedance r_0 in the saturation region has been made. In reality, however, this is not the real situation because r_0 increases with V_{DS}. In fact, an increment of the V_{DS} corresponds to a shift of the pinch-off point towards the source, resulting in a reduction of the rate with which the depletion region around the source becomes wider. Furthermore, as explained in the previous paragraph, also hot carriers play a role in the output conductance

determination.

The real consequences of all these effects can be however better understood through examples. Therefore, in the next section, a comparison between two different submicron technologies is performed, in order to highlight how the short channel effects influence the transistor behavior.

2.4 65nm vs 250nm

The comparison has been carried out between the 250nm CMOS technology and the 65nm CMOS technology, because the current CMS pixel detector uses a Front-End architecture based on the former, while the phase 2 upgrade detector will use the latter. Before starting, it has to be underlined that the supply voltage decreases from 2.5 V in the *250nm* to 1.2 V in the *65nm*. The analysis has been performed, in both cases, simply using a NMOS transistor biased as shown in Figure 2.9.

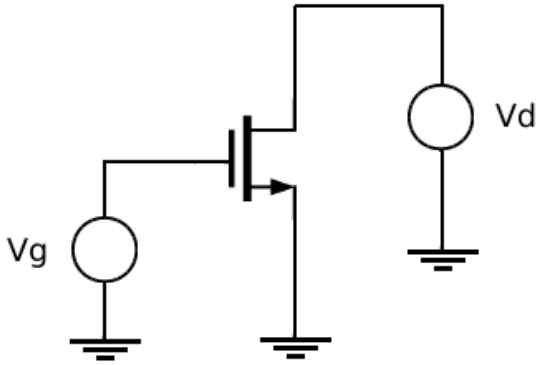


Figure 2.9: Simulation test bench

Threshold voltage

In Figure 2.10 also the *350nm* technology slope has been reported, in order to better understand the effects due to halo implants.

In the *350nm* technology halo doping is absent. Accordingly, V_{TH} increases with the channel length while in both *65nm* and *250nm* technologies halo doping is used, resulting in a decrease of the V_{TH} with the channel length. It is furthermore interesting to notice that in the *65nm* technology, unless the slope has the same behavior with respect to the *250nm*, the threshold value is a little higher, due to the different production processes.

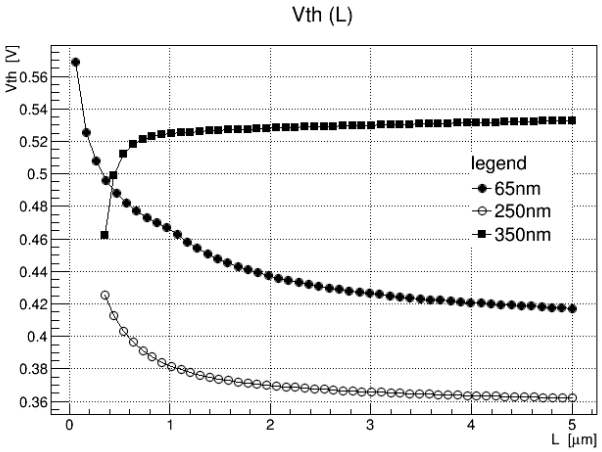


Figure 2.10: $V_{TH}(L)$ for different technologies

Output conductance and intrinsic gain

In all amplifiers the most important parameter is the gain. Then it is very useful to evaluate the intrinsic gain, i.e. the maximum gain achievable with a single transistor. In formula it can be expressed as follows:

$$A_0 = \frac{g_m}{g_{ds}} \tag{2.22}$$

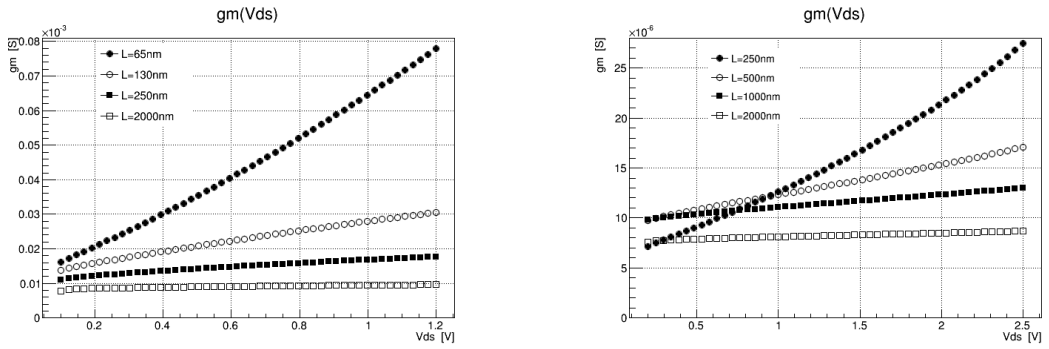
Subsequently, the first thing to do is the analysis of the g_m and g_{ds} behavior separately.

- **Transconductance** Focusing on the g_m behavior, shown in Figure 2.11, it is possible to see that at L_{min} the trend is quite different from the others in both technologies.

  Furthermore, g_m looks like to have a linear dependence on the V_{DS} value due to the channel length modulation effect. Choosing $L = 2\mu m$ in both the technologies it is possible to see that the trend is very similar. As a consequence, it is possible to say that there is not a big difference between the two technologies in terms of transconductance.

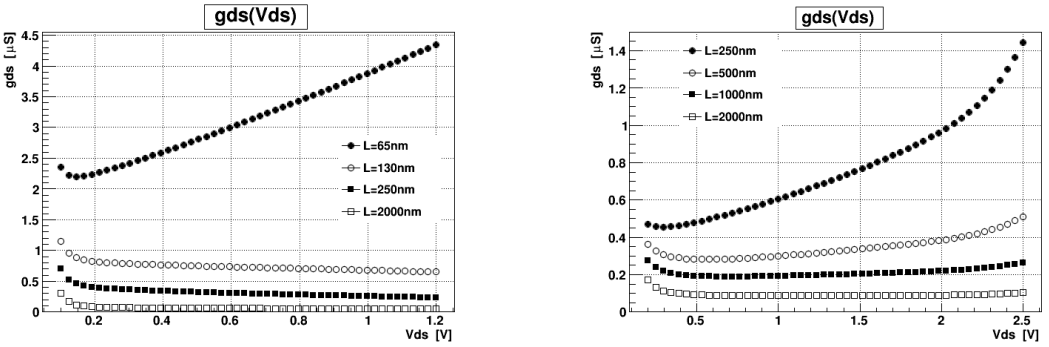
- **Output conductance** Let's go now to talk about g_{ds} . It is a key parameter because it is the reciprocal of the output resistance. A useful way to get some information about this parameter is to plot g_{ds} vs V_{DS} for different channel length values in both the technologies, as shown in Figure 2.12.

  Firstly, it has to be stressed that the L_{min} is not the right choice neither in *65nm* nor in *250nm*. In fact, the g_{ds} value is significantly higher, resulting,



(a) 65nm CMOS technology (b) 250nm CMOS technology

Figure 2.11: $g_m(V_{DS})$ for different technologies



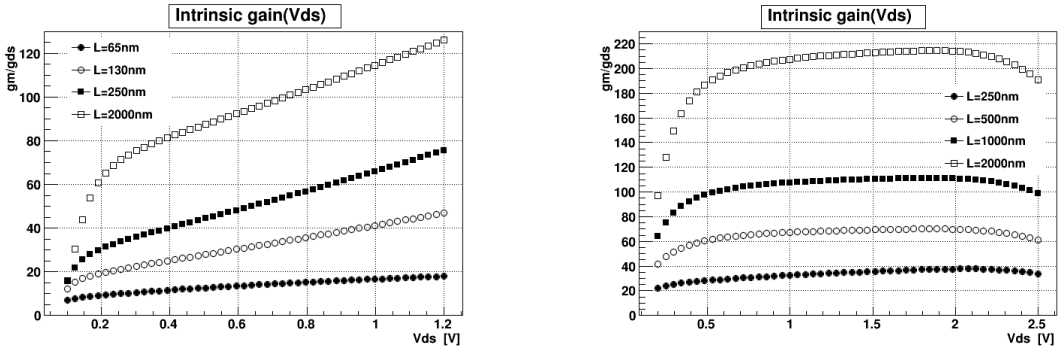
(a) 65nm CMOS technology (b) 250nm CMOS technology

Figure 2.12: $g_{ds}(V_{DS})$ for different technologies

as a consequence, in a smaller output resistance value. Furthermore, it is possible to see that the trend is different for high V_{DS} values, in fact in the *250nm* technology the output conductance value rises, effect which does not occur in the *65nm* . It is due to the impact ionization effect, which is probably better controlled in the *65nm* technology. Finally, choosing the same L value ($L = 2\mu m$) in both the technologies, it can be seen that the output conductance values is nearly the same in the two cases.

- **Intrinsic gain** Referring now to the intrinsic gain, the same kind of comparison has been performed.

  This comparison shows that the choice of small channel length values leads to little gain values. However, it is also clear that in the *65nm* technology



(a) 65nm CMOS technology

(b) 250nm CMOS technology

Figure 2.13: Intrinsic gain for different technologies

the intrinsic gain is reduced compared to the *250nm* case. In fact, choosing $L = 2\mu m$, the intrinsic gain is halved in the *65nm* technology.

Unity gain frequency

The unity gain frequency (f_T) is the frequency at which the transistor gain is 1. Accordingly it indicates the bandwidth of the single transistor. In order to find the f_T value, a plot of gain vs frequency has been realized, also using a $130nm$ technology to better explain the situation.

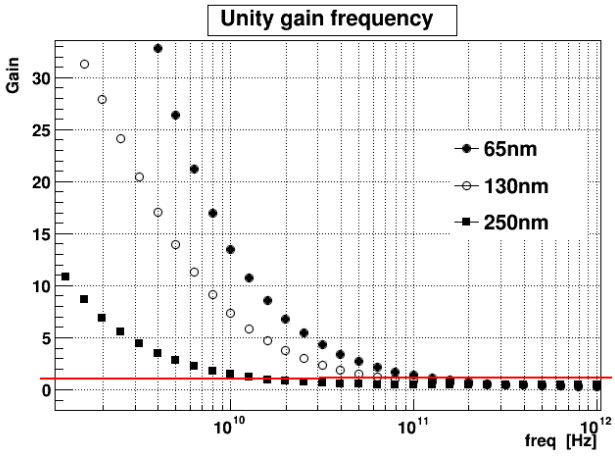


Figure 2.14: $Gain(f)$ for different technologies

Figure 2.14 shows that the unity gain value, represented by the red line, is reached at progressively higher frequencies as the technology node decreases. In the $65nm$

technology the unity gain value corresponds to a frequency of about 100 GHz, an order of magnitude higher than in the $250nm$ case, in other words the bandwidth is larger in $65nm$.

2.4.1 Summary

Putting things together, this comparison shows that the fundamental transistor parameters are not very different in the two technologies. The main difference which comes out is that with the decreasing of the technology node there is a progressive reduction of the intrinsic gain but, on the other hand, a widening of the bandwidth. The main advantages in using the $65nm$ is however that, as mentioned in Chapter 1, circuits occupy a smaller area and the overall power consumption is reduced, in particular in the digital part.

Chapter 3

Front-End amplifiers

In this chapter an overview of a typical electronics Front-End for silicon detectors will be given. In the second part the attention will be focused on noise in electronics systems. In this section the main noise contributions will be calculated, since their mathematical expressions are very important to understand the noise optimization procedure which will be described in the following chapter.

3.1 Front-End amplifier

In Figure 3.1 an overview of a Front-End amplifier is shown.

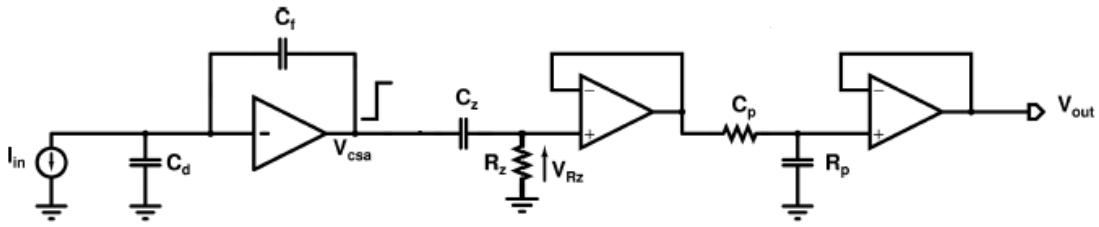


Figure 3.1: Front-End amplifier

The first stage, which directly receives the signal rising from the sensor, is the Charge Sensitive Amplifier (CSA). It is followed by a shaper, whose purpose is a further signal processing. Let's now analyze more in detail the behavior of the single stages.

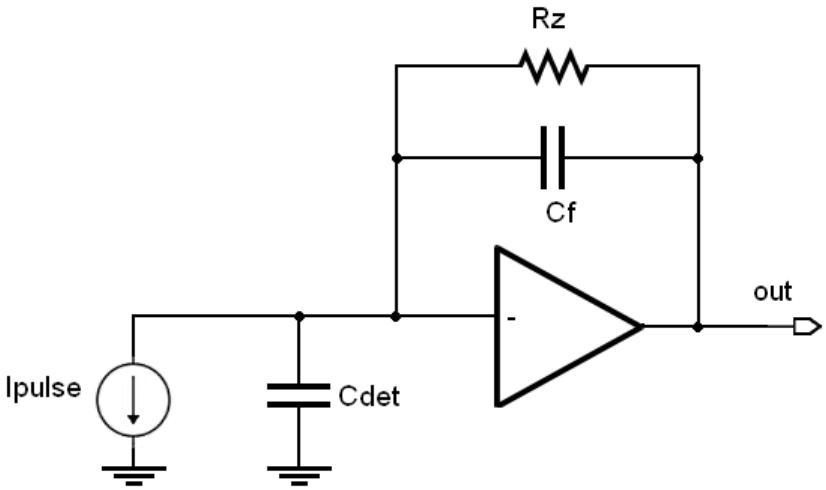


Figure 3.2: Scheme of a Charge Sensitive Amplifier

3.1.1 Charge Sensitive Amplifier

The Charge Sensitive Amplifier (CSA), shown in Figure 3.2, is the first element of the chain. The core component of the CSA is a high-gain amplifier with a feedback network implemented with a capacitor C_f. The resistor R_f is then added in order to provide a steady DC input level, as it is required by high-gain amplifiers. However, the resistor value is so high (about tens of GΩ) that it does not affect signal processing.
The following assumptions are useful to understand the behavior of an ideal CSA:

- The input signal is a δ-like pulse
- The core amplifier has infinite gain and infinite bandwidth

If these requirements are accomplished the CSA can be described as an ideal integrator, so:

$$V_{out} = \frac{1}{C_f} \int dt\ I_{in}(t) \tag{3.1}$$

With a δ-like input, ($I_{in}(t) = Q_{in}\ \delta(t)$) the previous relationship becomes:

$$V_{out} = \frac{Q_{in}}{C_f} u(t) \tag{3.2}$$

In the frequency domain, using the Laplace transform of u(t) which is $\frac{1}{s}$, the output signal can be rewritten:

$$V_{out} = \frac{Q_{in}}{C_f}\frac{1}{s} \tag{3.3}$$

In Figure 3.3 it is possible to see the output signal of an ideal CSA. When the input signal arrives, the CSA output reaches immediately its maximum. However, in real amplifiers, the rise time of the output signal is not negligible, and it is usually around some ns.

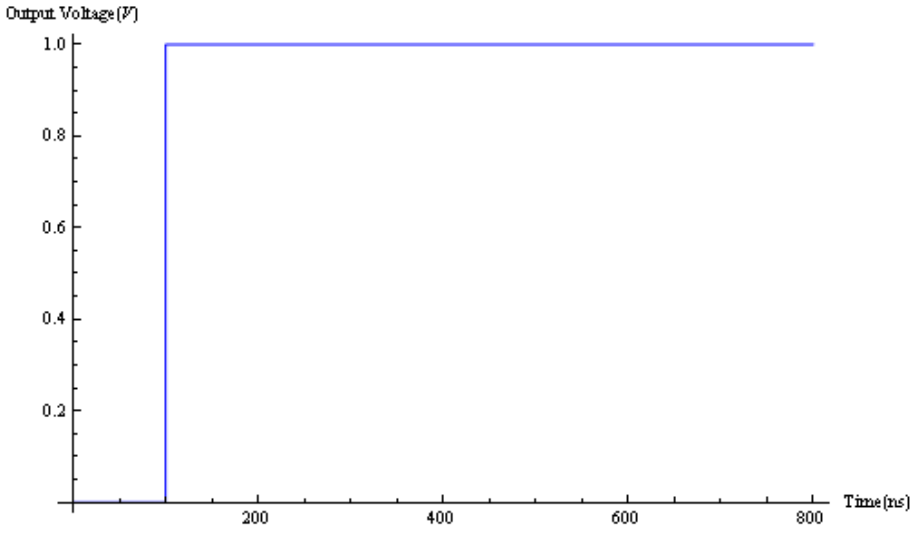


Figure 3.3: Output signal of an ideal CSA

3.1.2 CR-RC shaper

The CSA output signal has to be furthermore elaborated in order to facilitate the signal analysis performed in the successive electronics chain. This purpose is accomplished by the shaper stage. It can be implemented in many ways, but it is substantially composed of a series of differentiators and integrators. In Figure 3.4 a typical choice, the CR-RC shaper, is shown. It is a sequence of a high-pass filter, which is a differentiator, and a low-pass filter, which is an integrator and it is an appropriate choice to understand how the shapers behave.

In order to analyze the behavior of the stage, it is useful to write the transfer functions of the differentiator and the integrator in the frequency domain. The

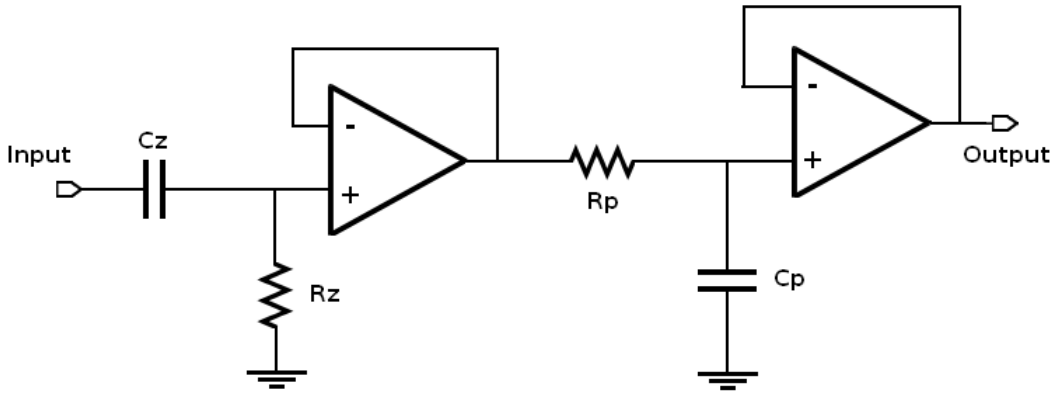


Figure 3.4: Scheme of a CR-RC shaper

high-pass filter transfer function is:

$$T(s) = \frac{sRC}{1 + sRC} \tag{3.4}$$

while the low-pass filter one is:

$$T(s) = \frac{1}{1 + sRC} \tag{3.5}$$

In order to understand step-by-step how the chain of CSA and shaper operates, it is firstly appropriate to consider the sequence of CSA and high-pass filter. It is now useful to write the output signal in the frequence domain using the Laplace transforms.

$$V_{R_z} = \frac{Q_{in}}{sC_f} \frac{sC_zR_z}{1 + sC_zR_z} = \frac{Q_{in}}{C_f} \frac{\tau_z}{1 + s\tau_z} \tag{3.6}$$

This expression shows that the pole introduced by the CSA is cancelled by the zero introduced by the high-pass filter. However, the expression in the time domain is convenient to understand the signal form shown in Figure 3.5, which shows that the output of the differentiator decreases with an exponential law with time constant τ_z.

$$V_{out,R_z} = \frac{Q_{in}}{C_f} e^{-\frac{t}{\tau_z}} \tag{3.7}$$

The purpose of the differentiator is thus to give rise to a signal which goes back to the baseline. It is now appropriate to underline that a very important information to be measured is the peak height, because it is proportional to the charge released

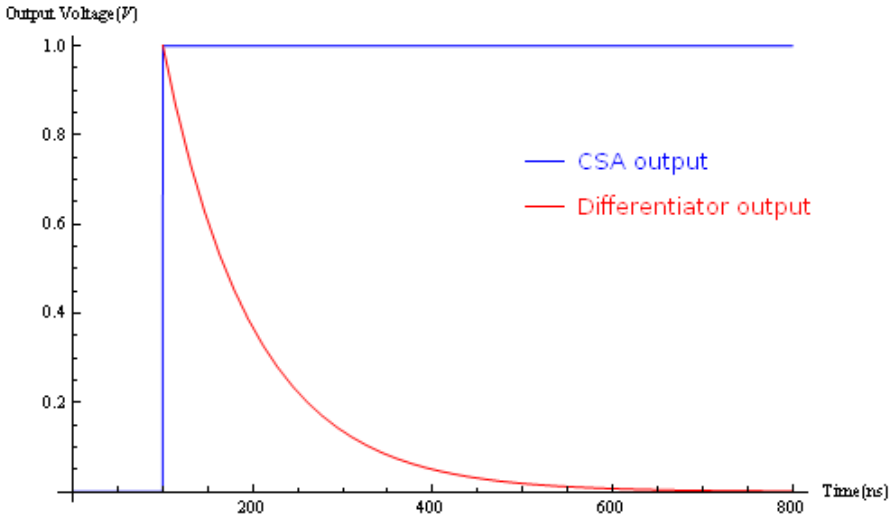


Figure 3.5: Output signal of the differentiator stage

in the sensor by the particles. But in this configuration the signal fall is very quick, requiring a circuit which makes a very precise measurement. For this reason it is convenient to diminish the variations around the maximum. This purpose is reached elaborating the signal in the integrator. The output signal of the full chain, calling τ_p the time constant of the integrator, is then given by the following relationship:

$$V_{out}(t) = \frac{Q_{in}}{C_f}\frac{\tau_z}{\tau_z - \tau_p}(e^{-\frac{t}{\tau_z}} - e^{-\frac{t}{\tau_p}}) \tag{3.8}$$

This expression is correct if the gain of differentiator and integrator is the same. If $\tau_z = \tau_p$ the previous expression is simplified:

$$V_{out}(t) = \frac{Q_{in}}{C_f}\left(\frac{t}{\tau}\right)e^{-\frac{t}{\tau}} \tag{3.9}$$

This particular case is very convenient. The return to the baseline is in fact dominated by the smaller time constant, so for example if the integration time constant is significantly greater than the other one, the integration of the signal will be very little. A good compromise between these different configurations is then the choice of equal time constants, usually made in CR-RC shaper.
Having made the time constants choice, it becomes interesting to determine the value of the peaking time, i.e. the time taken by the signal to reach its maximum. It can be obtained equalling to zero the first derivative of the output signal:

$$\frac{dV_{out}}{dt} = \frac{1}{\tau}e^{-\frac{t}{\tau}} - \frac{t}{\tau^2}e^{-\frac{t}{\tau}} = 0 \tag{3.10}$$

then

$$T_P = \tau \tag{3.11}$$

The maximum value of the signal at the shaper output is instead:

$$V_{out,max} = \frac{Q_{in}}{C_f}\frac{1}{e} \tag{3.12}$$

At the shaper output, therefore, there is a loss in the gain, which can be compensated by an increase of the gain of the following stages.

Until now an ideal Front-End description has been carried out, but obviously, it has to be implemented with real components: as a consequence in the circuit design it is necessary to take into account the intrinsic noise contribution which rise from the electronics chain. For this reason, the next part of the chapter is focused on this issue and on the calculation of the main contributions.

3.2 Noise in electronics systems

The noise analysis is a key element in the circuit design. In fact, especially in analog circuits for the readout of radiation sensors, the noise evaluation plays a key role, because it is crucial to have the best possible Signal-to-Noise Ratio (SNR). In other words, it is desirable to have the smaller possible influence of the noise on the signal form. Noise is a random process caused by the fluctuations of currents and voltages generated into the device [11]. As a consequence, it is not possible to predict the instantaneous noise value in the time domain, so the noise analysis is performed observing the noise for a long time developing a statistical model. It means that the predictable noise properties are average values. [1]. To each noise source is associated a power spectral density S_n^2, which indicates how much noise power is present at a given frequency. Furthermore, it is usual to call "white noise sources" the noise signals with a flat power spectral density, i.e. independent from the frequency.
Moreover, the electric charge is not a continuous quantity: the unity is the electron charge. For this reason, the noise is usually expressed in ENC(Equivalent Noise Charge), i.e. the number of electrons which must be collected in order to have a signal equivalent to the noise one.
In general, in the simulations the noise effects are measured at the output and then properly referred to the input. The power spectral density can be used to obtain the value of the square rms of the output noise generated by the input source by

multiplying it for the square of the noise source transfer function and then integrating on the frequency spectrum:

$$V_{n,out}^2 = \int_0^\infty df \; S_n^2 \, |T_n(j2\pi f)|^2 \tag{3.13}$$

The total output noise, considering uncorrelated sources, is then given by:

$$V_{n,out\ rms} = \sqrt{V_{n,out1}^2 + V_{n,out2}^2 + ... + V_{n,outn}^2} \tag{3.14}$$

The ENC of a single noise source is then given by:

$$ENC = \frac{1}{q_e} \frac{1}{A_0} V_{n,outn} \tag{3.15}$$

The total ENC instead is:

$$ENC_{tot} = \sqrt{ENC_1^2 + ENC_2^2 + ... + ENC_n^2} \; = \; \frac{1}{q_e} \frac{1}{A_0} V_{n,out\ rms} \tag{3.16}$$

q_e Electron charge expressed in Coulomb
A_0 Gain of the stage

3.2.1 Types of noise

The principal noise sources are the thermal and the flicker noise [12].

- *The thermal noise* is caused by the thermal excitation of charge carriers in a semiconductor. It has a white spectral density and is proportional to the absolute temperature.
- *The flicker noise* is present only with DC currents and originates from the traps in the semiconductor that hold the carriers which constitute the DC flow for some time before releasing them. It is often called "1/f noise" because its spectral density can be represented as a $1/f^\alpha$ function, $(0.8 < \alpha < 1.3)$, in which f is the frequency.

3.2.2 Noise in CMOS technologies

Also in MOS transistors the main noise sources are the thermal and the flicker noise. The MOS channel has a resistive behavior, consequently the contribution of thermal noise, as for resistors, is relevant. Flicker noise is instead due to the fact that in MOS transistor the current conduction is near to the surface, which behaves as traps, capturing and releasing charges. These statements have to be taken into account in the representation of the noise sources due to MOS transistors.

3.2.3 Noise in Front-End amplifiers

However, it is important to understand how to represent these noise sources in Front-End amplifiers. The most used method is the representation of the main noise contributions in terms of current of voltage sources. The distinction between these two cases is performed in the following way:

- **Series noise**: noise contribution represented by a voltage source put in series with the amplifier input
- **Parallel noise**: noise contribution represented by a current source placed in parallel with the amplifier input

Now a list of the main noise sources is given.

Series noise sources

- White noise due to resistors connected in series with the input:

$$V_{nR_s}^2 = 4kTR_s \tag{3.17}$$

  in which k is the Boltzmann constant and T is the absolute temperature.

- White noise due to the input transistor:

$$V_{nw1}^2 = 4kT\gamma\frac{1}{g_{m1}} \tag{3.18}$$

  in which g_{m1} is the transconductance of the input transistor and γ is the inversion factor introduced in 2.2.4

- White noise due to the current source:

$$V_{nw2}^2 = 4kT\gamma\frac{g_{m2}}{g_{m1}^2} \tag{3.19}$$

  in which g_{m1} and g_{m2} are the transconductances of the two transistor which form the current mirror, i.e. the implementation of the current source at the transistor level.

- Flicker noise intrinsically generated by the amplifier

$$V_{n1/f}^2 = \frac{K_f}{C_{ox}WL}\frac{1}{f} \tag{3.20}$$

  where K_f is a constant value once the gate length is fixed.

Parallel noise sources

All the following parallel noise sources have white spectral density. Some of them are internal to the Front-End amplifier:

- Noise due to current sources directly connected to the input

$$I_{ns}^2 = 4kT\gamma g_m \tag{3.21}$$

- Noise generated by the feedback resistor of the CSA:

$$I_{n,R}^2 = \frac{4kT}{R_f} \tag{3.22}$$

  This relationship shows that it is advantageous to have a large feedback resistor.

There is also another noise source, which does not depend on the electronics, but which has a great influence on the Front-End system. It is generated by the detector leakage current, which increases with the radiation damage. It is a parallel noise source:

$$I_{n,leak}^2 = 2qI_{leak} \tag{3.23}$$

Figure 3.6 shows how noise contributions are represented in a typical Front-End amplifier.

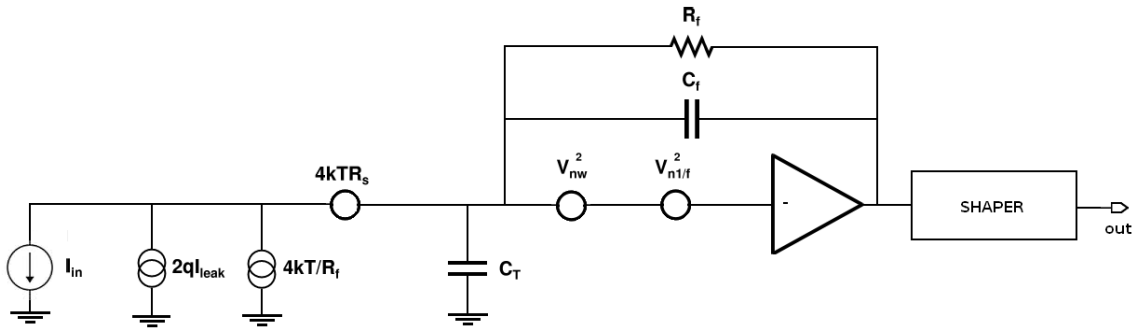


Figure 3.6: Noise sources

3.3 Noise calculations

At this point it is necessary to calculate the total ENC generated by the Front-End amplifier, following the method indicated by the equation 3.16. These calculations can be performed either in the time domain or in the frequency domain. In

this case the former is chosen , because it is particularly suitable in white parallel and white series noise calculations. Only the former will be presented in a complete way because the other case is very similar.

3.3.1 Parallel noise calculation

The starting assumption is that parallel noise can be modeled as a series of elementary pulses which randomly occur at the amplifier input. It means that these signal are processed by the electronics chain likewise the sensor signals, resulting in a perturbation of the output signal. It is hence crucial to understand how the parallel noise affects the measurement of the output signal performed at a generic time T_1. Remembering the equation 3.9, which expresses the amplifier response to a $\delta - like$ input at t=0, it is interesting to figure out which is the response to a noise pulse of charge q occurring at the measurement time T_1:

$$V_{outn}(t) = \frac{q}{C_f}\left(\frac{t-T_1}{\tau}\right)e^{-\frac{t-T_1}{\tau}} \tag{3.24}$$

from which:

$$V_{outn}(T_1) = 0 \tag{3.25}$$

In other words, it means that noise pulses occurring at the moment in which the measurement is performed or afterwards, do not influence the output signal. Therefore, from now on only pulses occurring before the measurement are considered.
If the noise pulse arrives at the input β seconds before the measurement, the equation 3.24 becomes:

$$V_{outn}(t) = \frac{q}{C_f}\left(\frac{t-(T_1-\beta)}{\tau}\right)e^{-\frac{t-(T_1-\beta)}{\tau}} \tag{3.26}$$

then the output voltage value at the measurement time is:

$$V_{outn}(T_1) = \frac{q}{C_f}\frac{\beta}{\tau}e^{-\frac{\beta}{\tau}} \tag{3.27}$$

A particular case is when $\beta = T_P = \tau$, where T_P is the shaper peaking time. The equation3.27 becomes:

$$V_{outn}(T_1) = \frac{q}{C_f}\frac{1}{e} \tag{3.28}$$

i.e. the noise perturbation reaches its maximum. It is now clear that it would be useful to build a quantity indicating the importance of the parallel noise contribution. In order to get it the equation 3.26 can be rewritten observing that $\beta = T_1 - t$:

$$V_{outn}(t) = \frac{q}{C_f}\left(\frac{T_1-t}{\tau}\right)e^{-\frac{T_1-t}{\tau}} = qW(t) \tag{3.29}$$

where $W(t)$ is called "Weighting function". An alternative expression is the following, in which $W(t)$ in divided by its maximum, in order to have an input referred noise contribution expressed in units of charge.

$$W_N(t) = e\left(\frac{T_1 - t}{\tau}\right) e^{-\frac{T_1 - t}{\tau}} \tag{3.30}$$

Let's now use the weighting function to calculate the parallel noise contribution, making the assumption that the noise events at the input follow a Poisson distribution, in which average value and variance have the same value. If n_{noise} is the rate of noise events, the average value of noise events occurring in a time dt is given by $n_{noise}dt$. Referring all to the input, each noise event gives a contribution equal to $qW_N(t)$. It means that the variance originating by $n_{noise}dt$ pulses is $n_{noise}q^2W_N^2(t)dt$. The total variance of the input charge, expressed in $[Coulomb]^2$, is then given by the following expression:

$$\sigma_q^2 = n_{noise}q^2 \int_{-\infty}^{+\infty} dt(W_N^2(t)) \tag{3.31}$$

Using that

$$n_{noise}q^2 = qI \tag{3.32}$$

where I is the average current related to the process, it is possible to rewrite the equation 3.31:

$$\sigma_q^2 = 2qI\ \frac{1}{2}\int_{-\infty}^{+\infty} dt(W_N^2(t)) \tag{3.33}$$

Furthermore, remembering the expression of the weighting function 3.30 and that pulse events occurring after the measurement are negligible, it is possible to evaluate the integral:

$$\int_{-\infty}^{+\infty} dt(W_N^2(t)) = \int_{-\infty}^{T_1} dt(W_N^2(t)) = e^2\frac{\tau}{4} = e^2\frac{T_P}{4} \tag{3.34}$$

The final expression of the variance is then:

$$\sigma_q^2 = 2qI\frac{e^2}{8}T_P \tag{3.35}$$

The parallel noise value expressed in ENC is given by:

$$ENC_{parallel} = \sqrt{\frac{\sigma_q^2}{q^2}} \tag{3.36}$$

This relationship is very important, because it shows that the ENC of parallel noise

- depends on the square-root of the peaking time
- is independent from the total capacitance

Series noise and flicker noise

The weighting function formalism allows also to derive the ENC for the series noise contribution using a similar procedure. The final result is:

$$ENC_{series} = \frac{V_{nw}}{q} C_T \sqrt{\frac{e^2}{8} \frac{1}{T_P}} \tag{3.37}$$

This quantity is

- linearly dependent on the total capacitance
- dependent on the reciprocal of the square-root of peaking time

As a consequence, unlike the parallel noise, the series noise tends to decrease if the peaking time is greater. It means that it is possible to find a value of the peaking time in which the best compromise between the two contributions can be achieved. It also shows a direct dependence on the total capacitance.

The contribution of the flicker noise is given by the following relationship:

$$ENC_{1/f} = (C_D + C_G) \sqrt{\frac{K_f}{C_{ox} W L} N_{1/f}} \tag{3.38}$$

so it is

- linearly dependent on the total capacitance
- independent from the peaking time

In this context it is useful to verify how the different ENC contributions change with the peaking time. The situation is shown in Figure 3.7.

It is appropriate to underline that, in the architectures studied in the following chapters, the chosen peaking time is 12.5 ns. In fact, in the LHC every 25ns a collision occurs, so it can give rise to an event which has to be processed by the electronics chain. This value is marked by the red line in Figure 3.7 in order to emphasize that in this case the prevailing contribution is the series noise, element which will be taken into account in the following chapters.

Since the expressions of the main noise contribution are now known, the next step is to understand how to perform the noise optimization, i.e. the minimization of the ENC value for a Front-End amplifier. This topic is discussed in chapter 4.

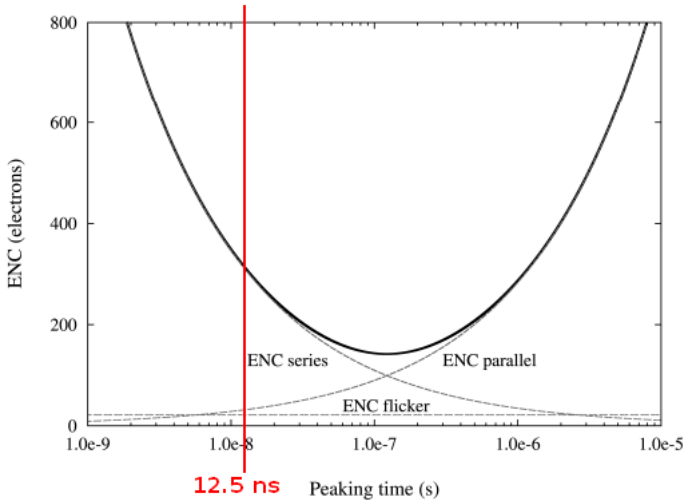


Figure 3.7: ENC vs peaking time

Chapter 4

Noise optimization of the Charge Sensitive Amplifier

This chapter discusses of the noise optimization methods in Front-End amplifiers. The first part of the chapter is about the optimization procedure, which can be dependent on the operating region of the input transistor. In fact, the classical technique described in the first paragraph is effective only in strong inversion. Then a more modern method based on the EKV model is illustrated, as it is suitable in all the operating regions, including the weak inversion region, very common in deep submicron CMOS technologies.
Subsequently, in the second part the Charge Sensitive Amplifier studied in this work is presented with a particular focus on the noise optimization of the stage.

Before starting, it is necessary to stress that, for the time being, the parallel noise contribution has not been considered, since it is quite small with the chosen peaking time, as explained in Chapter 3.

4.1 Noise optimization of a Charge Sensitive Amplifier

In order to have the best signal-to-noise ratio from the CSA it is necessary to minimize the ENC figure. This goal can be achieved through an optimized sizing of the input transistor, which is the main noise source in the chosen CSA architecture. The approach to the problem is different if the transistor operates in strong inversion or not. In the first case the optimization rule is quite simple, but in modern technologies CMOS transistors works in weak or moderate inversion due to the current densities used in Front-End circuits. This situation forces the development of a new model suitable to describe the noise effects in all the operating regions.

4.1.1 Transistor in strong inversion

Thermal noise

The first thing to do is to write the equation describing the thermal and flicker noise contribution. For the former, which is usually the most relevant one, the expression is the following:

$$ENC_{th}^2 = 4kTn\gamma\alpha_w \frac{(C_D + C_G)^2}{g_m(C_G)} \frac{N_{th}}{T_p} \tag{4.1}$$

with:

k	Boltzmann constant
T	Absolute temperature
n	Slope factor($\simeq 1.5$)
α_w	Excess noise factor
γ	Inversion factor
N_{th}	Shaper noise index for thermal noise
T_P	Peaking time

In Equation 4.1 the dependence of g_m on the gate capacitance is highlighted to stress that a compromise between two opposite trends has to be found. In order to identify the minimum ENC value it is then necessary to equal to zero the first derivative of Equation 4.1, which gives rise to the following relationship:

$$2g_m = (C_D + C_G)\frac{dg_m}{dC_G} \tag{4.2}$$

Equation 4.2 can be further manipulated. Using the fact that the transistor is in strong inversion, it is possible to write again the g_m expression found in 2.2.4.

$$g_m = \mu C_{ox}\frac{W}{L}(V_{GS} - V_{TH}) = \sqrt{2\mu C_{ox}\frac{W}{L}I_{DS}} \tag{4.3}$$

Using now that, in first approximation, the input transistor gate capacitance is:

$$C_G = C_{ox}WL \tag{4.4}$$

Equation 4.3 can be written in the following way:

$$g_m = \sqrt{2\mu C_G \frac{I_{DS}}{L^2}} \tag{4.5}$$

Differentiating this espression with respect to C_G and inserting the result in the 4.2 we find:

$$\sqrt{2\mu C_G \frac{I_{DS}}{L^2}} = (C_D + C_G)\frac{\mu\frac{I_{DS}}{L^2}}{\sqrt{2\mu C_G \frac{I_{DS}}{L^2}}} \tag{4.6}$$

that, with some calculations, becomes:

$$4C_G = C_D + C_G \tag{4.7}$$

The final condition which minimizes the thermal noise contribution of the input transistor is then:

$$C_G = \frac{1}{3}C_D \tag{4.8}$$

Flicker noise

Referring now to the flicker noise, the ENC expression is:

$$ENC_{1/f}^2 = \frac{K_f}{C_{ox}WL}C_T^2 N_{1/f} \tag{4.9}$$

with:

W,L	channel width and length
C_{ox}	gate capacitance per unit area
K_f	flicker noise coefficient
$N_{1/f}$	shaper noise index for flicker noise
C_T	total input capacitance

C_T can be expressed as the sum of the detector capacitance C_D and the gate capacitance $C_G = C_{ox}WL$:

$$ENC_{1/f}^2 = K_f\frac{(C_D + C_G)^2}{C_G}N_{1/f} \tag{4.10}$$

Also in this case there is a minimum value which identifies the best choice, found differentiating with respect to C_G. This time the condition is:

$$C_D = C_G \tag{4.11}$$

This value is quite different from the best choice in terms of thermal noise, so the final choice of the C_G will depend on the specific situation: the element to be judged is which of these two noise contributions is prevalent.

4.1.2 Transistor in moderate or weak inversion

A more frequent case in deep submicron technologies is when the input transistor works in moderate ($0.1 < I_C < 10$) or weak inversion ($I_C < 0.1$). In this configuration the previous technique is not valid anymore.
An alternative way is to determine a model which gives the chance of expressing

relevant quantities as functions of the inversion coefficient. It avoids, as a consequence, the dependence of the noise optimization on the transistor working region. In other words, this different approach leads to a method which is valid in all operating regions. An important example is the following one, based on the EKV (Enz-Krummenacher-Vittoz) model [13], in which the transconductance is given by:

$$g_m = \frac{I_D}{n\phi_T} \frac{1}{\sqrt{I_C + 0.5\sqrt{I_C} + 1}} \tag{4.12}$$

while the gate capacitance, given by:

$$C_G = C_{GB} + C_{GS} \tag{4.13}$$

can be modeled with:

$$C_G = \frac{n - (1+x)/3}{n} C_{ox} W L = C(x(I_C)) C_{ox} W L \tag{4.14}$$

The quantity x, in turn, is defined in the following way:

$$x = \frac{(\sqrt{I_C + 0.25} + 0.5) + 1}{(\sqrt{I_C + 0.25} + 0.5)^2} \tag{4.15}$$

In addition, it is appropriate to recall that the source and the drain electrodes extend themselves a little underneath the gate, giving rise to a further contribution to the gate capacitance. This is called "overlap capacitance" and its value is around $1fF/\mu m^2$. The total gate capacitance becomes:

$$C_G = C(x(I_C)) C_{ox} W L + 2C_{ov} W = C_{GW}(I_C, L) W \tag{4.16}$$

Furthermore, since it will be used in the following paragraphs, it is appropriate to recall the inversion factor (γ) expression, already defined in chapter 2:

$$\gamma = \frac{1}{2} + \frac{1}{6} \frac{I_C}{I_C + 1} \tag{4.17}$$

An interesting question to answer is how much these expressions are compatible with the values of transconductance and gate capacitance given by the simulator. A simple way to verify it is to use a diode-connected transistor as shown in the Figure 4.1.

The idea is to change the current flowing from drain to source in order to change the inversion coefficient, fixing all the other parameters, such as the aspect ratio of the transistor. In this way it is possible to compare the values of g_m and C_G given by the formulas with the simulation ones. The results of this work are shown in Figures 4.2 and 4.3.

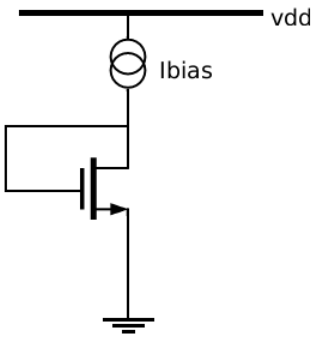


Figure 4.1: NMOS diode-connected

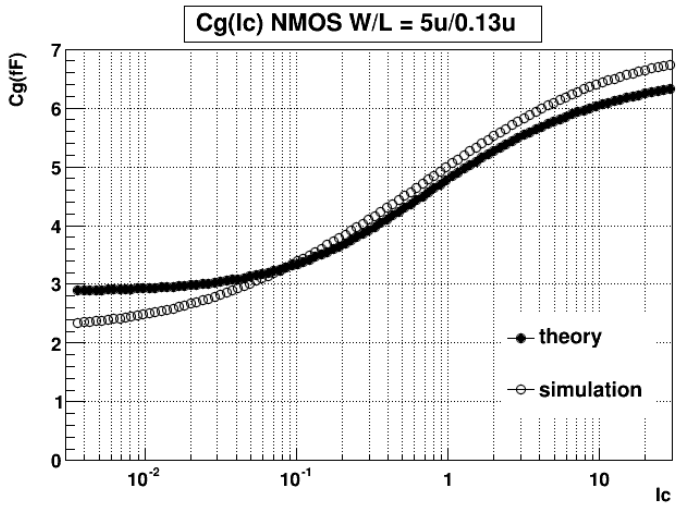


Figure 4.2: C_g vs I_C for a diode-connected NMOS

As shown in the figures the matching between model and simulation is quite good, especially for the transconductance. However, in both cases in a large spectrum of I_C the difference between the two trends is below the 10%.

In order to be completely sure about the use of this model the same comparison has been performed with a PMOS diode-connected, as shown in Figures 4.4 and 4.5.

However the results are very similar to the previous ones. These confirmations on the reliability of the EKV model allows to use it to make further considerations about noise.

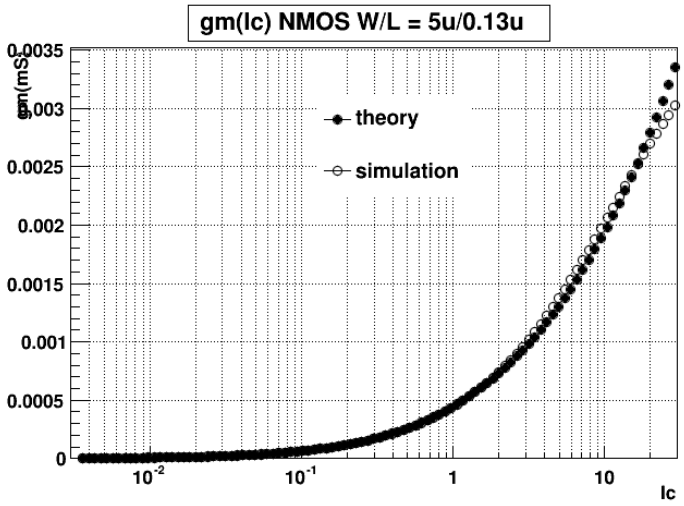


Figure 4.3: g_m vs I_C for a diode-connected NMOS

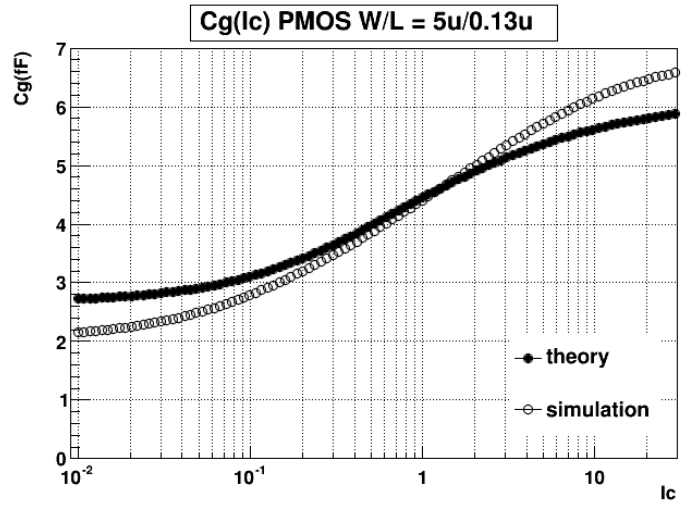


Figure 4.4: C_g vs I_C for a diode-connected PMOS

The overall ENC equation can be rewritten taking into account that now the most part of the quantities depend on the inversion coefficient, so the expression is valid in all the transistor operating regions.

$$ENC^2 = (C_D + C_{GW}(I_C,L)W)^2 \left[\frac{4kTn\gamma\alpha_w}{g_m(I_C)}\frac{N_{th}}{T_p} + \frac{K_f}{C_{ox}WL}N_{1/f}\right] \tag{4.18}$$

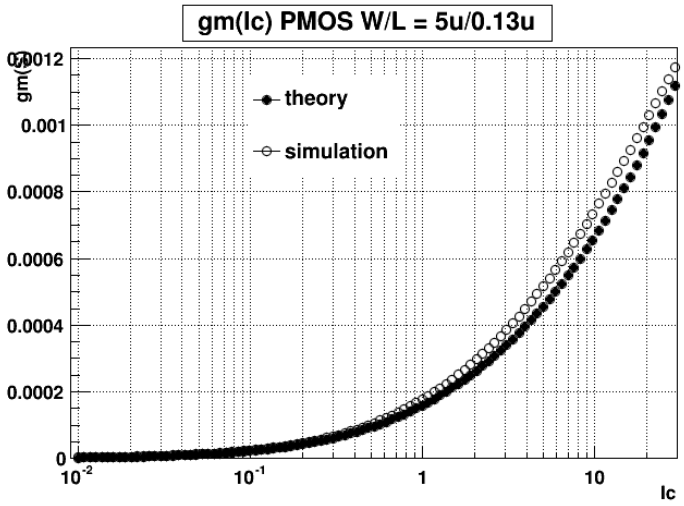


Figure 4.5: g_m vs I_C for a diode-connected NMOS

At this point the detector capacitance, the peaking time and the noise indexes are known. The designer has thus to determine the best transistor sizing in order to minimize the ENC contribution, taking into account the power consumption. However, in order to better understand the procedure, the analysis performed for the architecture studied in this work is explained in the following paragraph.

4.2 Studied architecture

4.2.1 Sensor model

Before starting with the simulations, it is appropriate to explain how the sensor behavior is modeled in the Front-End analysis. The representation of the sensor as a circuit element is presented in Figure 4.6.

The model consists of three components:

- I_{pulse} A pulse generator which simulates the current pulse produced by a particle into the sensor
- C_{det} It is the sensor capacitance, i.e. the the capacitance of the single pixel cell. Planar silicon sensors have a capacitance of around $100 fF$, which becomes around $400 fF$ for 3D silicon sensors.
- I_{leak} A DC current generator which simulates the leakage current of the detector.

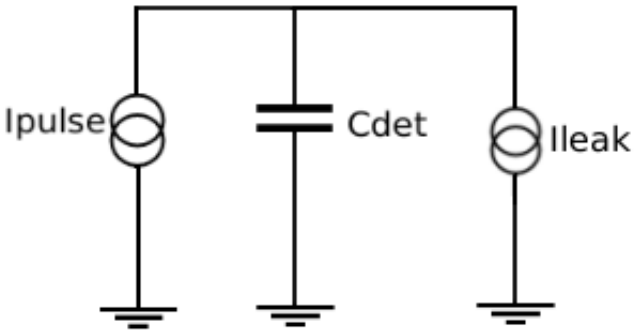


Figure 4.6: Sensor model

The detector leakage current tends to increase with the radiation damage of the detector. At the beginning it is around $10nA$, but during the years it can rise until around $100nA$. This current can strongly influence the behavior of the Front-End amplifier, so it needs a specific compensation architecture which will be discussed at the end of Chapter 5. Therefore, in these first analyses the leakage current has been neglected.

4.2.2 Charge Sensitive Amplifier architecture

In Figure 4.7, the chosen architecture for the study of the noise is presented. The first stage is composed of a common source amplifier with a cascoded NMOS and the current source implemented with a PMOS current mirror. The $1pF$ capacitance is inserted in order to cut the noise contributions rising from M3. The second stage, instead, is a source follower. In Table 4.1 the transistor sizes and the other circuit parameters are listed. It is important to observe that the widths of the input transistor and of the cascode one are not indicated because their values will be found at the end of the noise optimization procedure.

Taking into account the EKV model statements, the idea is to plot the ENC behavior in terms of different quantities, such as W and L of the input transistor, the current flowing into the transistor and the detector capacitance, with the purpose of determine the best conditions which minimize the noise.

4.2.3 Channel width analysis

The first parameter which has been taken into account is the input transistor channel width (W).

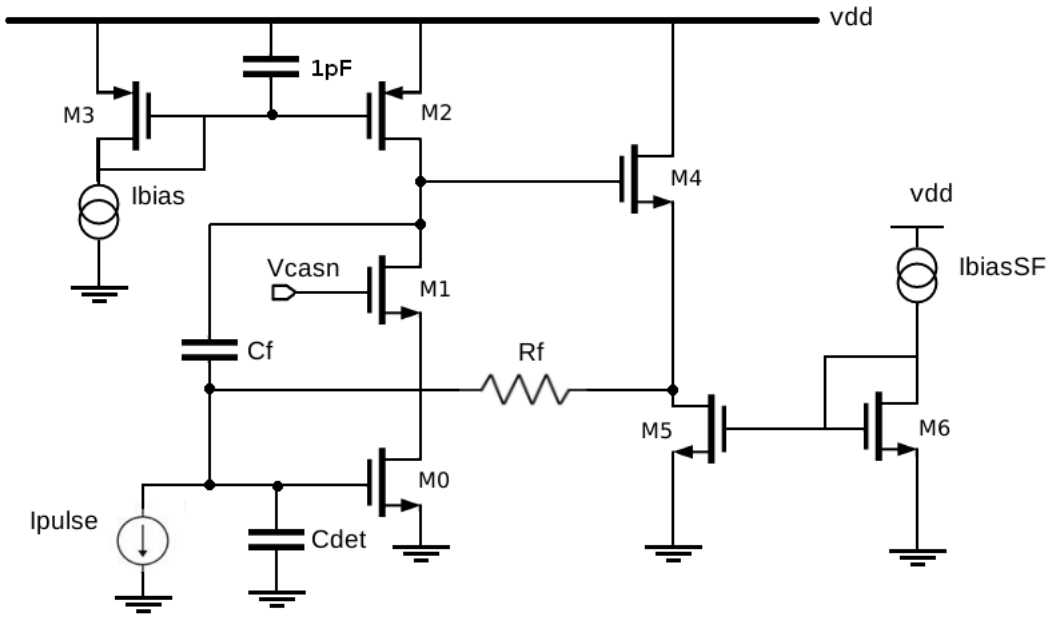


Figure 4.7: CSA at the transistor level

(a) Transistors aspect ratio

Transistor	$W(\mu m)$	$L(\mu m)$
$M0$	W	0.13
$M1$	W/4	0.13
$M2$	1	2
$M3$	0.5	2
$M4$	600	0.13
$M5$	0.5	1
$M6$	0.25	1

(b) Current and capacitance values

I_{bias}	1 μ A
$I_{bias,SF}$	350nA
$C_{detector}$	100fF
$C_{feedback}$	5fF
V_{casn}	600mV

Table 4.1: Values of the circuit parameters

The analysis, presented in Figure 4.8, shows that with small W values the ENC is high, then it significantly decreases when W increases. Considering channel widths greater than 6-8 μm, however, the ENC levels off. Further increasing the W values would only result in a greater area occupation and in a reduced speed, so the chosen W for the input transistor is $W = 8\mu m$.

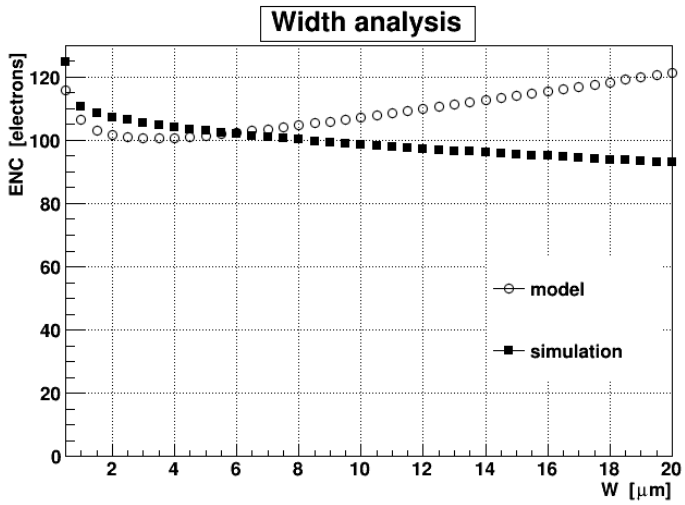


Figure 4.8: Channel width analysis

4.2.4 Channel length analysis

It is instead interesting to evaluate the channel length contribution to the noise. The analysis is again performed changing the W values, but for different cases, each of them corresponding to a different L value.

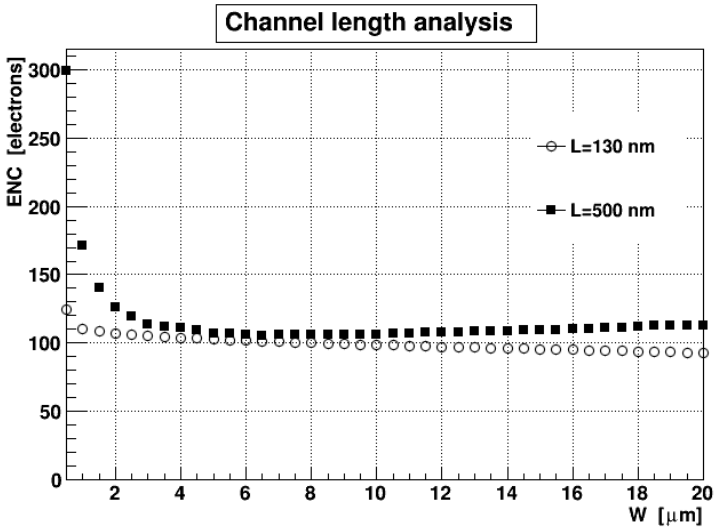


Figure 4.9: Channel width analysis for different channel lengths

Figure 4.9 shows that the ENC is significantly higher when the channel length

is 500nm compared to 130nm. Therefore, it is convenient to choose the smallest possible length. Since L_{min} has to be avoided due to the degradation of the transistor performances such as the output conductance, as shown in 2.4, the most appropriate choice is then $L = 130nm$.

4.2.5 ENC vs current analysis

Another key parameter is the current. In fact, the lower is the current, the smaller is the power consumption of the stage. The analysis has been performed with the transistor size fixed according to the previous analyses ($W = 8\mu m$ and $L = 130nm$).

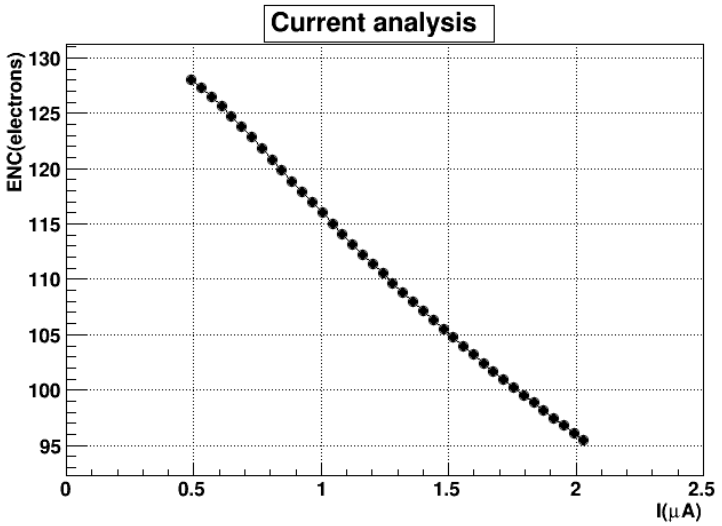


Figure 4.10: Current analysis

Figure 4.10 shows that decreasing the current the ENC significantly rises. A trade-off between these two requirements is then needed. As explained in chapter 1, the CMS phase 2 upgrade ROC has to meet several requirements. From the power consumption point of view the chip the power budget has to be below $0.4W/cm^2$. Assuming that the pixel dimension will be $100 \times 25\mu m$ and that the analog part and the digital part equally divide the area, all the analog part of a single pixel must not exceed a power consumption of around

$$P = 0.2\frac{W}{cm^2} \times 100\ 10^{-4}cm \times 25\ 10^{-4}cm = 5\mu W. \tag{4.19}$$

As a consequence, this single stage has to stay below it. A good compromise between noise and power consumption requirements is the choice of a current of $1\mu A$. In fact, since the supply voltage in this technology is $1.2V$, the power consumption is $1.2\mu W$.

4.2.6 Sensor capacitance analysis

Lastly, also the sensor capacitance contribution has to be evaluated. Since the peaking time is around 10 ns, the main noise contribution is the series noise, which is linearly dependent on the total input capacitance, given by the sum of the input transistor gate capacitance and the sensor capacitance. A planar silicon sensor has a capacitance of 100 fF, while a 3D silicon sensor has 400 fF. Both these values are significantly higher than the gate capacitance, which, for the chosen values of W and L is around $5fF$. It means that the dominant contribution is given by the sensor capacitance. The analysis is performed changing the W value, in order to verify if the previous considerations about the channel width are confirmed, but each curve corresponds to a different capacitance value.

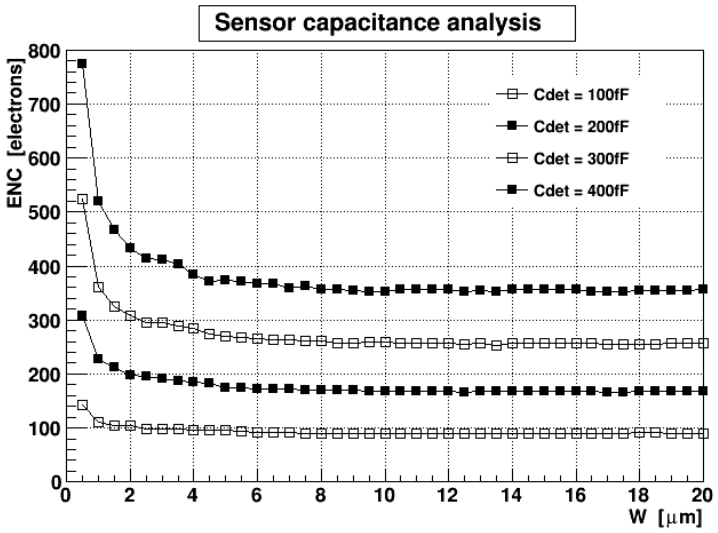


Figure 4.11: Channel width analysis for different sensor capacitances

Figure 4.11 shows that the choice of $W = 8\mu m$ remains good for different sensor capacitances, but the ENC increases significantly with the capacitance value. It becomes interesting to fix a value of W to control the trend of the ENC versus the sensor capacitance. Figure 4.12, realized with $W = 8\mu m$, confirms that the ENC rises linearly with the detector capacitance, so if a 3D silicon sensor is chosen it has to be taken into account that the ENC value will be 4 times higher than with a planar silicon sensor with the same electronics.

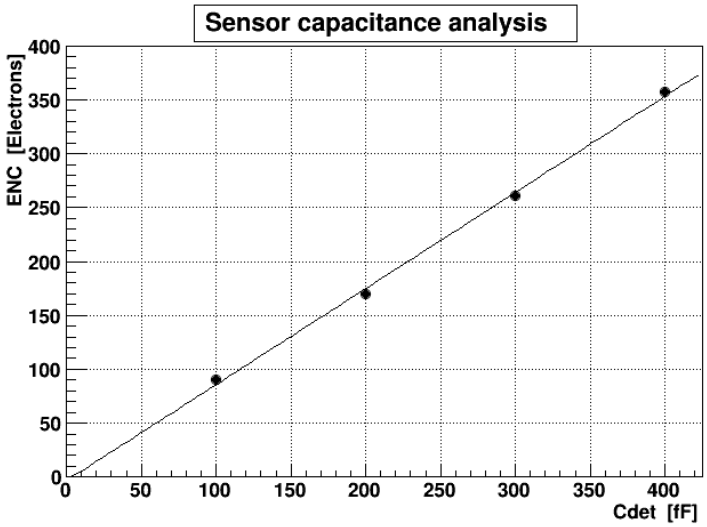


Figure 4.12: ENC vs sensor capacitance, $W = 8\mu m$

4.2.7 Summary

The simulation results determine the input transistor sizing. In fact, taking into account the considerations made in the previous paragraph, in order to minimize the ENC the most appropriate choice is a transistor with aspect ratio $\frac{W}{L} = \frac{8\mu}{0.13\mu}$. Accordingly, from now on the input transistor size will not be changed anymore. Furthermore, the current flowing in this stage will be always kept around $1\mu A$ in order to keep the power consumption under control.

Chapter 5

Results of simulations

This chapter is about the next steps of the analysis. In fact, the circuit shown in Chapter 4 is not the final version of the Front-End amplifier. In the previous architecture the feedback resistor was implemented with a passive component, which is not the right choice for Integrated Circuits, because it is very difficult to fabricate resistors with well-controlled values or with small physical sizes. The feedback network has to be implemented with active components, i.e. with transistors behaving as resistors. The first part of this chapter is dedicated to this topic, showing a comparison between two possible solutions in terms of noise, power consumption and mismatch performances. This work is performed using the results of the CAD simulations realized with the Cadence Virtuoso Simulator. The following part of the chapter describes instead possible solutions to take care of the detector leakage current, which until now has been neglected. In addition, at the end of the chapter some prospects about the development of a time-variant filter are given.

5.1 Analysis

Before starting, it is useful to enumerate some quantities which play a key role in the analysis of the preamplifier performance. The typical form of the output signal is shown in Figure 5.1.
The main features are shown in figure, but it is appropriate to better explain their role:

- **Baseline** It is the DC voltage value of the output, i.e. the value of the output before the start of the signal, and the value to which the output goes back after the signal reaches its peak

- **Peaking time** It is the time required for the output signal to go to the maximum value starting from the baseline. It is important to underline that

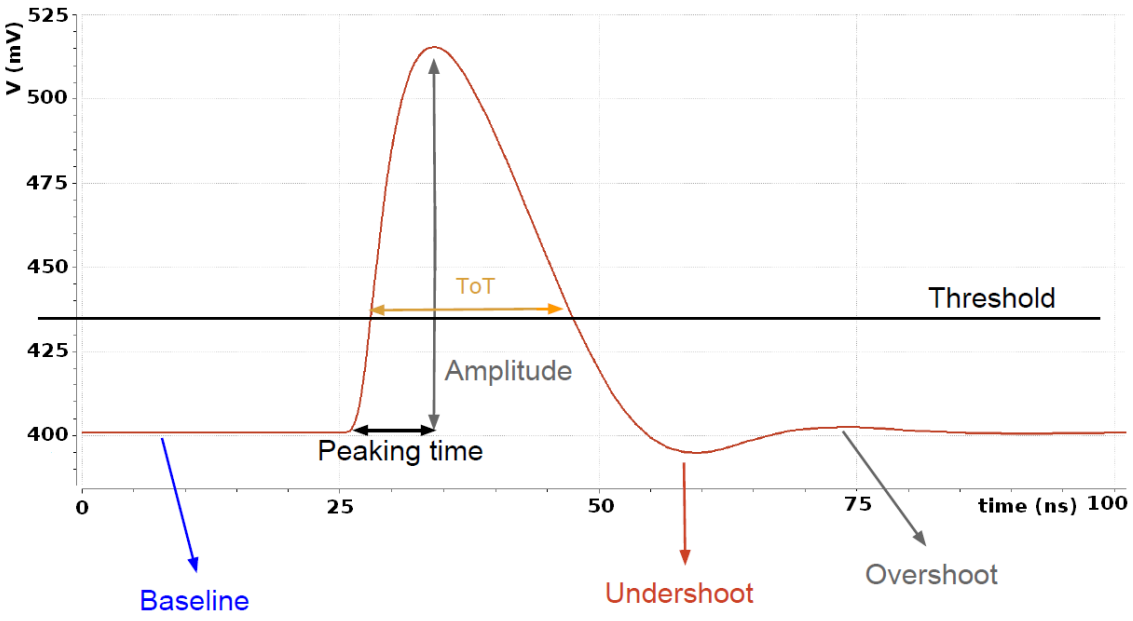


Figure 5.1: Typical output signal form

the output signal can be a little delayed with respect to the input signal, usually of around 1-2 ns. As a consequence, the output signal starts after the input activation.

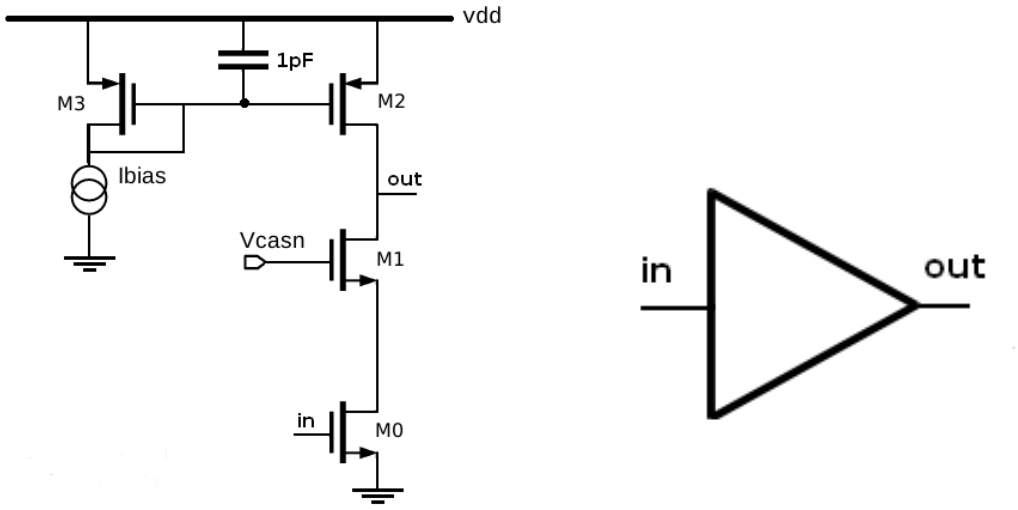
- **Amplitude** The amplitude of the signal is the difference between the voltage value of the peak and the voltage value of the baseline.

- **Threshold and ToT** It is common to choose a reference voltage value in order to measure the "Time over Threshold", i.e. the time in which the signal stays over the threshold value.

- **Undershoot and overshoot** Once the maximum is reached, an ideal signal simply returns to the baseline value. However, in real circuits some oscillations can occur, giving rise to the undershoot and overshoot of the output signals. It has to be underlined that undershoots and overshoots are indesiderable effects, because they can overlap with a second signal given by another particle, leading to a distortion of the latter.

5.2 Feedback network implementation

The implementation of a feedback network with active components can be made in some different ways. In this chapter two possible choices are described and compared in order to point out which of them is the most suitable for the upgrade project. It has to be underlined that in this chapter the first branch of the architecture presented in chapter 4, shown in Figure 5.2(a), is inserted in the symbol shown in Figure 5.2(b).



(a) Core amplifier Schematic

(b) Core amplifier symbol

Figure 5.2: Core amplifier representations

5.2.1 First architecture

The first architecture is shown in Figure 5.3. Before starting the analysis of this architecture it is appropriate to fix the circuit parameters. The table 5.1 shows the different choices for the transistors aspect ratio, the bias currents and voltages and the capacitance values.

This architecture is unipolar, but it is possible to make it almost bipolar, i.e. it can process both positive and negative signals. If the capacitor C1 is not connected, a positive signal is well processed, while more issues rise with the negative one, as it is shown in Figures 5.4 and 5.5.

It happens because the positive signal leads to a reduction of the V_{GS} of M3, which goes into the linear region. In order to fix this voltage, the capacitor C1 is

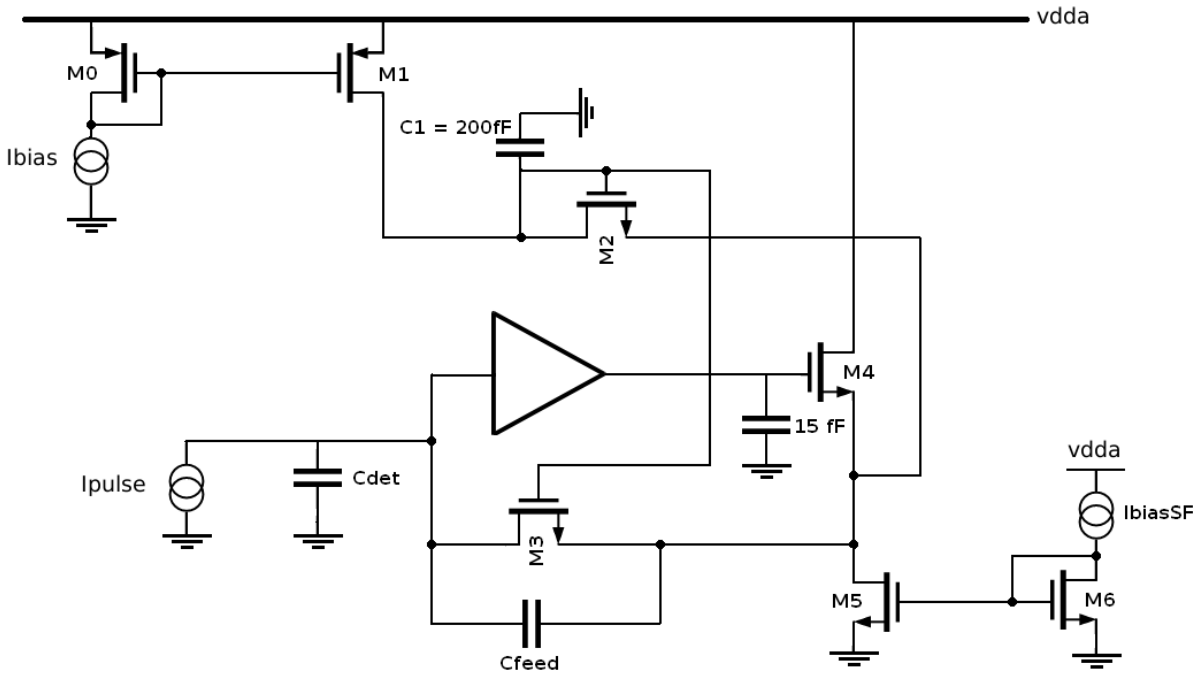


Figure 5.3: Schematic of the first architecture

(a) Transistors aspect ratio

Transistor	$W(\mu m)$	$L(\mu m)$
$M0$	0.25	1
$M1$	0.25	1
$M2$	0.5	0.5
$M3$	0.5	0.5
$M4$	1.5	0.3
$M5$	0.5	1
$M6$	0.25	1

(b) Current and capacitance values

I_{bias}	50nA
$I_{bias,SF}$	350nA
$C_{detector}$	100fF
$C_{feedback}$	5fF
V_{casn}	600mV

Table 5.1: Values of the circuit parameters

introduced. The first step is the choice of the value of C1. In fact, it has to be quite large so that the V_{GS} value remains almost steady, but not too much because it would be difficult to implement it in an Integrated Circuit. A simulation of the output signal with different capacitance values is then carried out to determine the best one, as shown in Figure 5.6.

$200fF$ is a good compromise. In fact, with higher capacitances the shape is only

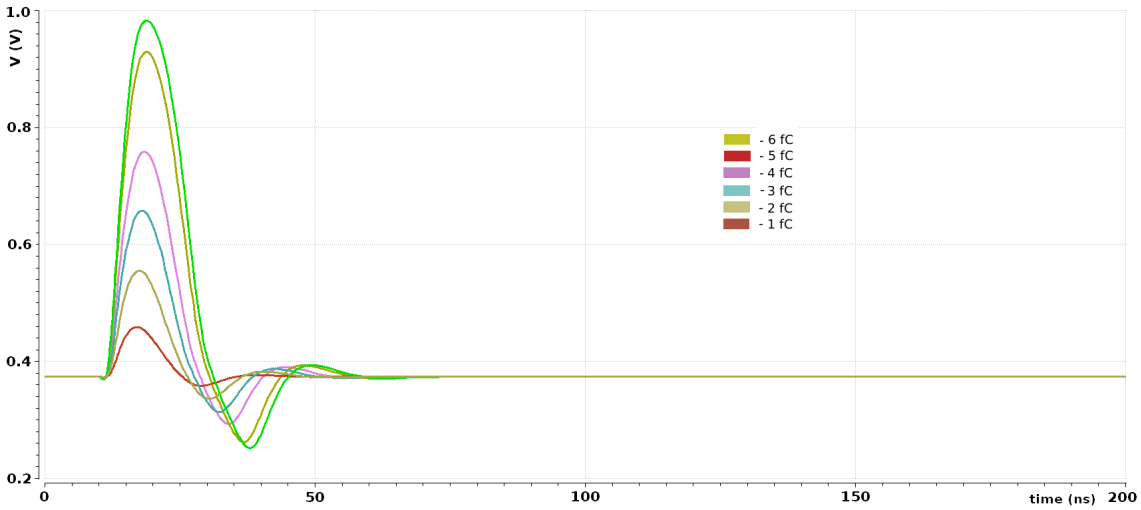


Figure 5.4: Output signal with a negative input without C1

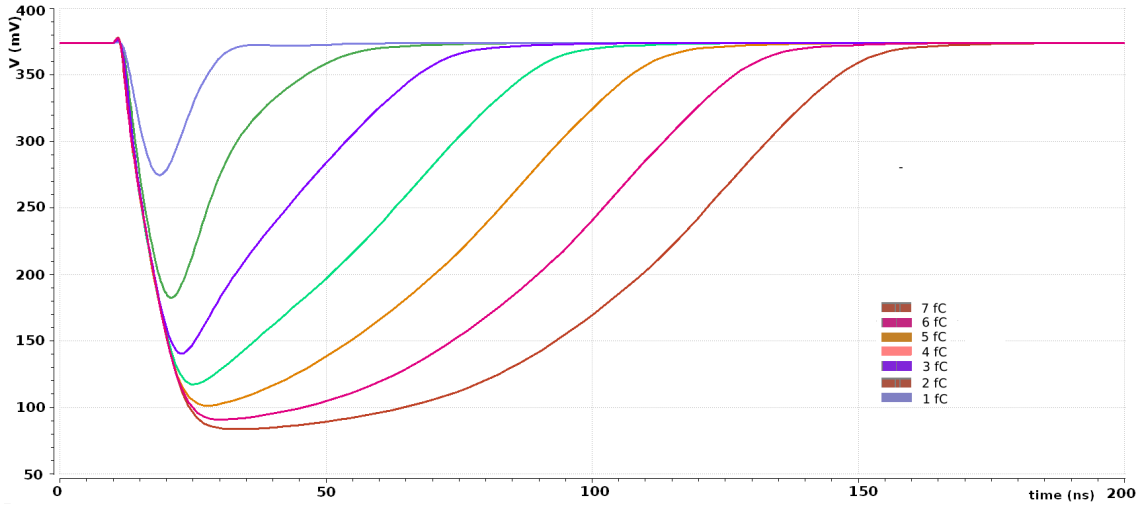


Figure 5.5: Output signal with a positive input without C1

a little better and, on the other hand, a capacitance of $200fF$ can be quite easily used in an Integrated Circuit. This figure also shows that, introducing the capacitor C1, the output shape is actually improved. Accordingly, it is possible to say that the previous issue is solved, but a new one rises, this time for positive signals, as shown in Figures 5.7 and 5.8. Therefore, the configuration which makes this architecture bipolar requires that the capacitor is connected only when the input signal is negative. This goal is achieved by connecting C1 with a CMOS switch.

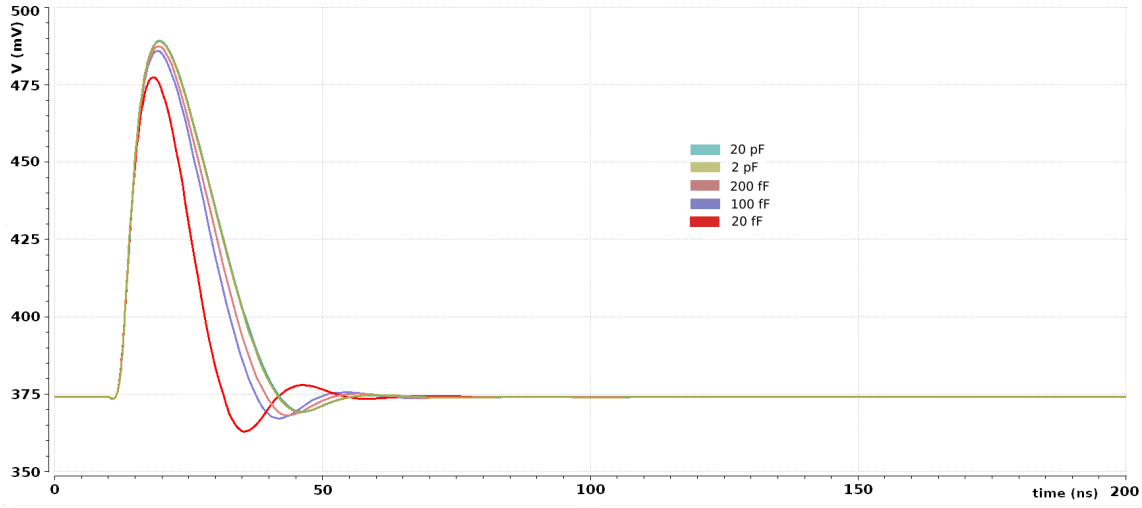


Figure 5.6: Output signal with a negative input for different C1 values

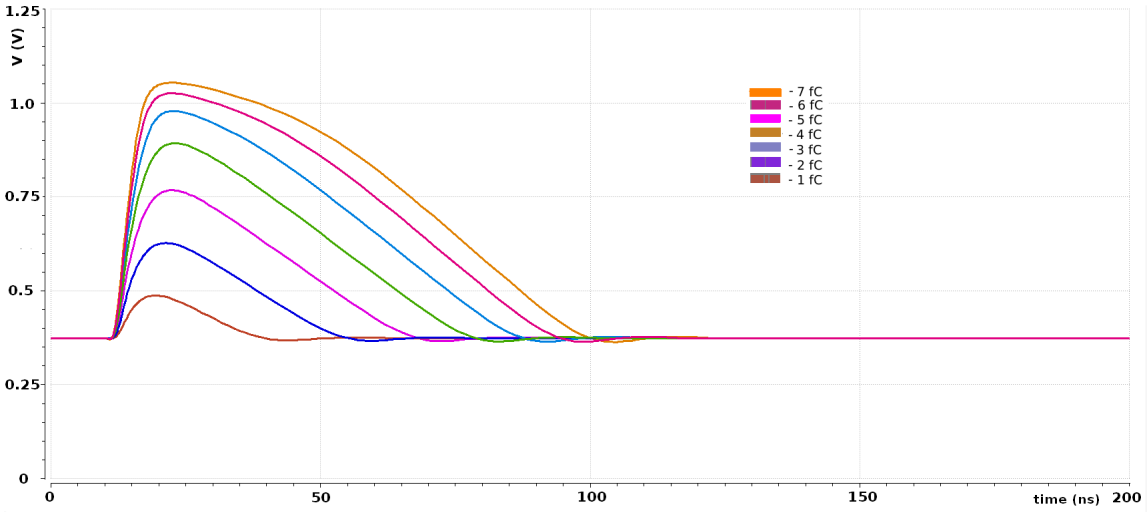


Figure 5.7: Output signal with a negative input

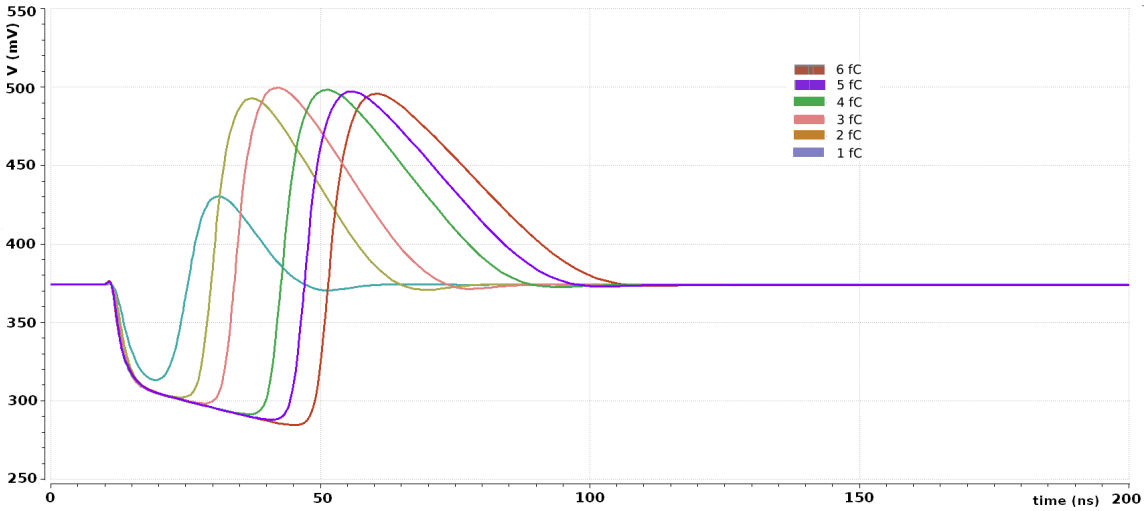


Figure 5.8: Output signal with a positive input

Power consumption estimation

The first parameter which has to be checked is the power consumption of the stage. The CAD simulator allows to simulate the power consumption of each device included in the circuit. With the previous choices about the bias current, the total static power consumption of this architecture is around $2\mu W$. A further optimization of the stage would be useful, because according to the system specifications, the power consumption of the complete analog chain has to be below $5\mu W$. Nevertheless, this is a good first step since this value is already under the limit.

Gain linearity

It is very important to check if the gain linearly depends on the input charge. It has to be verified in order to establish a linear relationship between the energy released by the charged particle into the sensor and the gain of the amplifier. Since in the upgraded sensor a MIP (Minimum Ionizing Particle) will produce on average a signal of $5 - 10Ke^{-}$, it is crucial to have linearity in this range of input charge.
In this case a transient time simulation is needed. This type of CAD simulation shows the time evolution of the output signal for a chosen time interval. Once the signal form is generated, it is necessary to measure the signal amplitude, which corresponds to the maximum gain of the stage. This procedure is repeated for different values of the input charge, i.e. for different input signals.
The results of the simulation are shown in Figure 5.9. The plot shows that the linearity is well verified not only in the range of the MIP charge, but in a much greater interval.

Transient noise analysis

The next step is the transient noise analysis, i.e. the noise effects are shown on the time evolution of the signal. In other words, it is always a transient time simulation with the addition of the noise effects, resulting in a fluctuation of baseline and signal values. This simulation is very important in order to fix a correct value of the threshold. It makes necessary the inclusion of new simulation options for the noise, as shown in the Table 5.2.
The frequency interval is specified because the noise is frequency dependent. T_{min},

$Noise\ Freq_{max}$	10 GHz
$Noise\ Freq_{min}$	100 mHz
$Noise\ T_{min}$	10 ps

Table 5.2: Transient noise analysis parameters

instead, is the minimum time step in the simulation. It can not be too small,

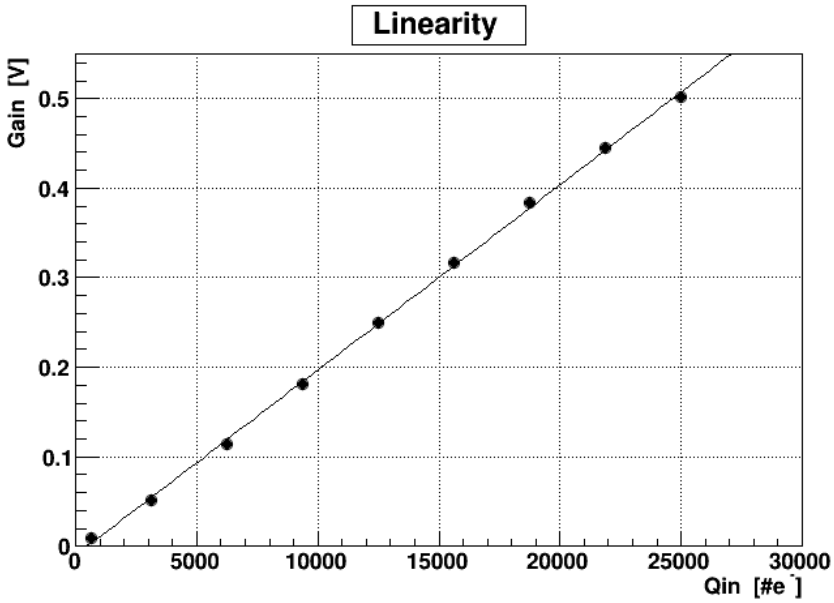


Figure 5.9: Gain vs input charge

otherwise the simulation becomes too slow. This simulation has been realized setting an input signal of $1fC$, i.e. $6250e^-$. This procedure is repeated 20 times. The superposition of these plots is shown in Figure 5.10. At this point it is crucial to find the distribution of the baseline values due to noise. The procedure is the following: once the results of the simulations are obtained, a histogram of the baseline values in the area marked in red in Figure 5.10 is realized, then the sigma of the distribution is taken. An acceptable choice for the threshold is hence:

$$V_{Threshold} = 5 \times V_{noise,rms} \tag{5.1}$$

The baseline fluctuations due to noise, in fact, follow a gaussian distribution, so $5\sigma \simeq 10^{-6}$. In other words, it means that only one noise event every million is over the threshold, leading to a wrong detection.

In Figure 5.11, instead, the histogram which comes from the baseline fluctuation analysis is shown. As already done before, it is more useful to convert the values expressed in Volts in values expressed in units of the electron charge. It is firstly necessary to convert the charge from Coulomb to units of electron charge in the following way:

$$q[electrons] = \frac{q[Coulomb]}{1.6 \times 10^{-19} C} \tag{5.2}$$

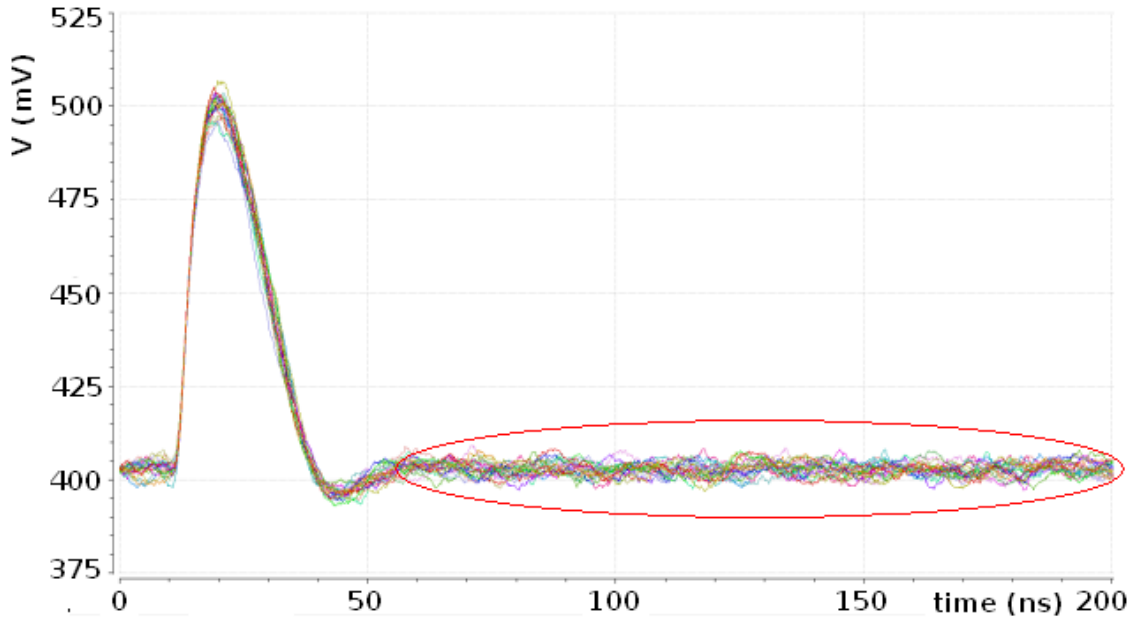


Figure 5.10: Transient noise analysis

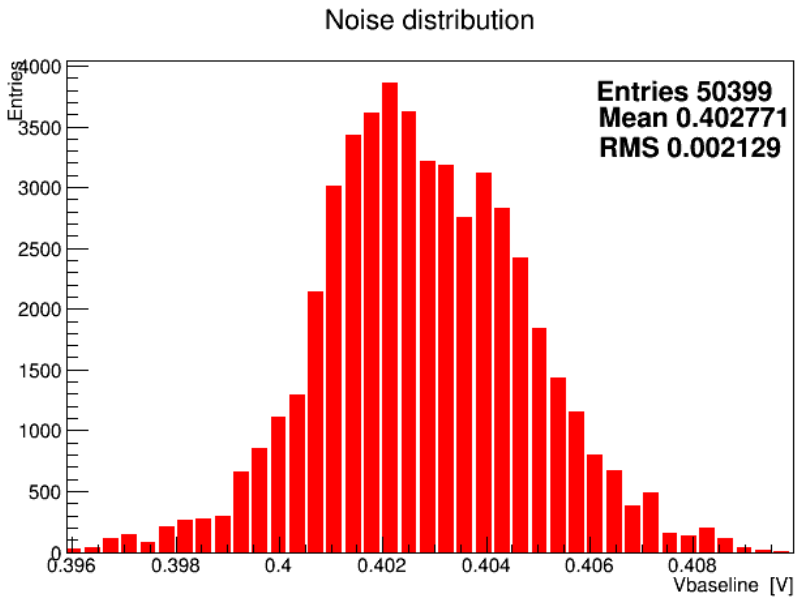


Figure 5.11: Histogram of the baseline fluctuations due to noise

As a consequence the RMS noise value becomes:

$$RMS[electrons] = \frac{q[electrons] \times RMS[V]}{Gain[V]} \tag{5.3}$$

in which the value is divided by the gain because it has to be an input-referred value, so the contribution given by the gain of the amplifier has to be deleted. So in this case, when the input charge is 1 fC, i.e. 6250 electrons, the gain is 110 mV/fC and then:

$$RMS[electrons] = \frac{6250 \times 2.1}{110} \simeq 120e^{-} \tag{5.4}$$

As a consequence, the threshold value chosen for the remaining analysis on this architecture is:

$$Threshold = 5 \times RMS[electrons] \simeq 600e^{-} \tag{5.5}$$

Mismatch simulation

A key check which has to be made is the mismatch simulation. The problem is that two nominally-identical circuits in practice have a different behavior due to the fluctuations of their transistors parameters caused by uncertainties in each step of the manufacturing process [1]. The random parameter fluctuations between transistors located on the same chip are called "mismatch variations". In addition, also systematic fluctuations between different silicon wafers, called "process variations", play a relevant role. The process variations have the same impact on all the transistors of a single wafer. Hereinafter both these effects will be grouped under the name "mismatch variations", also because the following simulations have been performed taking into account both these effects.
Therefore these aspects have to be carefully considered in the circuit design, because the produced transistor may not have the expected properties. The main reasons of these phenomena are:

- Fluctuations of the dopant concentrations
- Fluctuations of the oxide thickness

With regard to this issue, the Monte Carlo analysis, in which it is possible to simulate for a number of times the transistor of a specific technology, is very important. The CAD Monte Carlo simulation works in the following way: the mismatch model libraries, given by the foundry and which express the expected distribution of the transistor parameters, are included in the simulator. For each simulation run, the CAD randomly extracts these parameters from the previously described distribution and then simulates the circuit behavior. So, running the Monte Carlo simulation

several times, it is possible to get the distribution of quantities such as the ToT and the amplitude, due to the mismatch effects.
In order to better understand the situation, the behavior of a simple NMOS has been analyzed.

The idea of the simulation is the following. The W and the L are fixed. In this

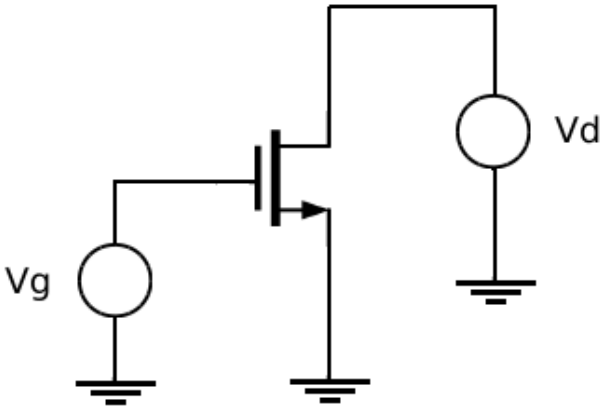


Figure 5.12: NMOS transistor

case the choice is $W = 2\mu m$ and $L = 1\mu m$. At this point a Monte Carlo simulation of 250 events is launched. It means that the simulator extracts each time the values of the transistor parameters from their distribution. In other words, the behavior of 250 apparently identical transistors is simulated, in order to understand how much mismatch effects change their properties. As an example, the Figure 5.13 shows the distribution of the NMOS threshold voltage. In addition, it is important to underline that the fluctuation of the electrical quantities of a transistor is dependent on the channel area. This statement can be easily verified repeating the previous procedure for some different values of W and L. In each case a histogram similar to Figure 5.13 has been carried-out, obtaining the corresponding value of $\sigma_{V_{TH}}$. The plot of the values of $\sigma_{V_{TH}}$ vs the square root of the transistor area is presented in Figure 5.14.

It is possible to see that the fluctuations of the threshold voltage significantly decrease with the increase of the square root of the transistor area, meaning that the mismatch effects are very relevant for transistors with an area lower than $1\mu m$. In first approximation it is possible to say that:

$$\sigma_{V_{TH}} \propto \frac{1}{\sqrt{WL}} \tag{5.6}$$

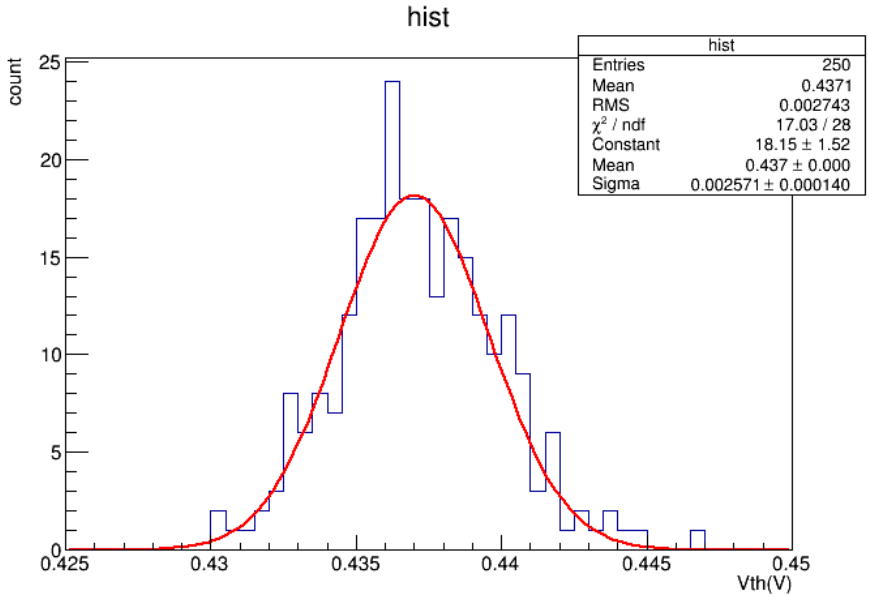


Figure 5.13: V_{TH} distribution for a NMOS with $\frac{W}{L} = \frac{2\mu m}{1\mu m}$

In other words, the smaller the transistor is, the higher the mismatch effects are. It is then clear that in a circuit with a lot of transistors it is crucial to monitor the mismatch contributions, in order to avoid that two practical implementations of the same architecture have very different behaviors.

Returning now to the preamplifier architecture, it is appropriate to highlight that in this case the principle of the Monte Carlo mismatch analysis is the same, only taking into account that there are a lot of transistors. A simulation of 250 events has been launched. The simulation gives then rise to a family of plots of the output signal, one for each event. Figure 5.15 shows the results obtained with the transistor sizing described in Table 5.1. As one can see, the mismatch effects have a strong influence on the output signal, especially on the signal duration and on the baseline value. In particular, in this case the most relevant issue is the variation of the signal duration. The first thing to do is to identify the main mismatch sources. An effective technique is to simulate the mismatch effects only for some transistors.

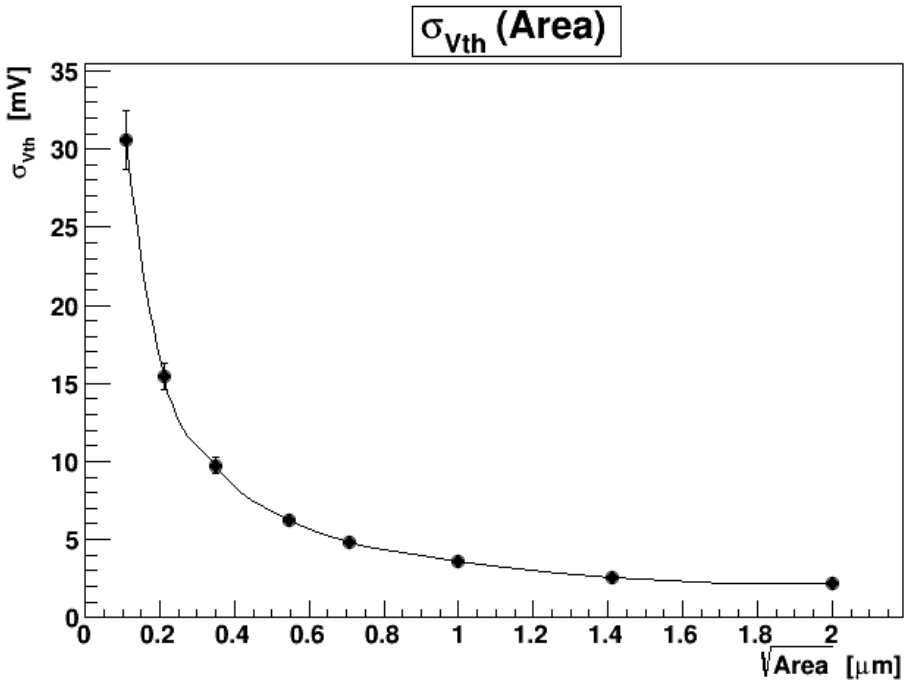


Figure 5.14: σ_{Vth} vs. $\sqrt{Area}$

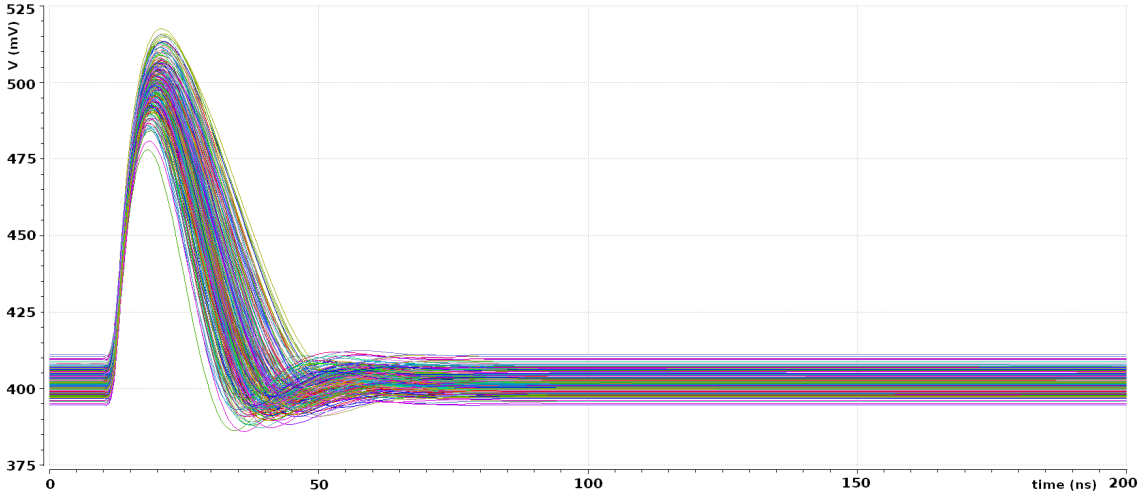


Figure 5.15: Mismatch analysis

- **M2 and M3** Figure 5.16 shows that these two transistors give a relevant contribution on the signal form, but not on the baseline value.

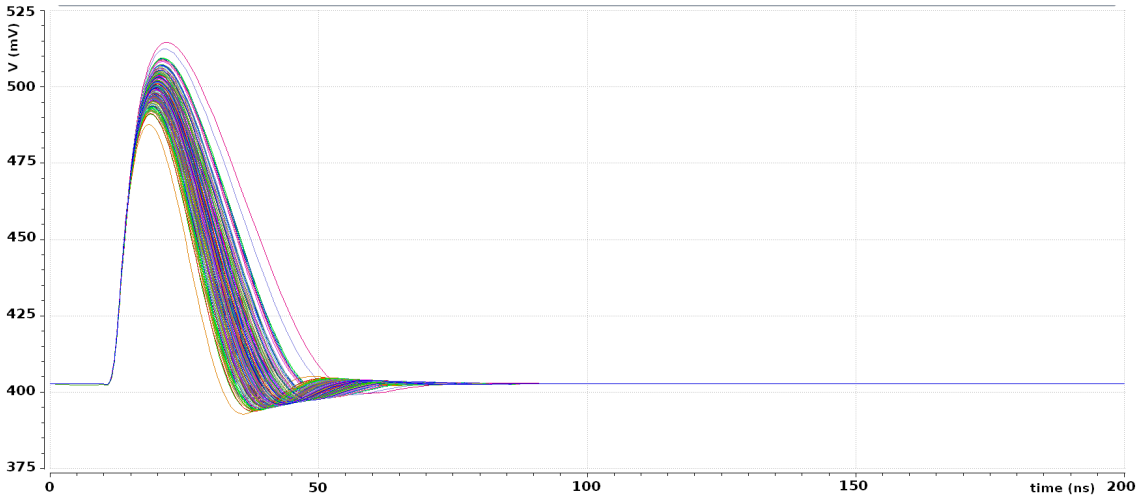


Figure 5.16: Mismatch analysis with M2 and M3 only

- **M0 and M1** As reported in Figure 5.17, the current mirror composed of M0 and M1 gives a contribution very similar to the previous one. As a consequence, the optimization of these two quantities will be performed in the same way.

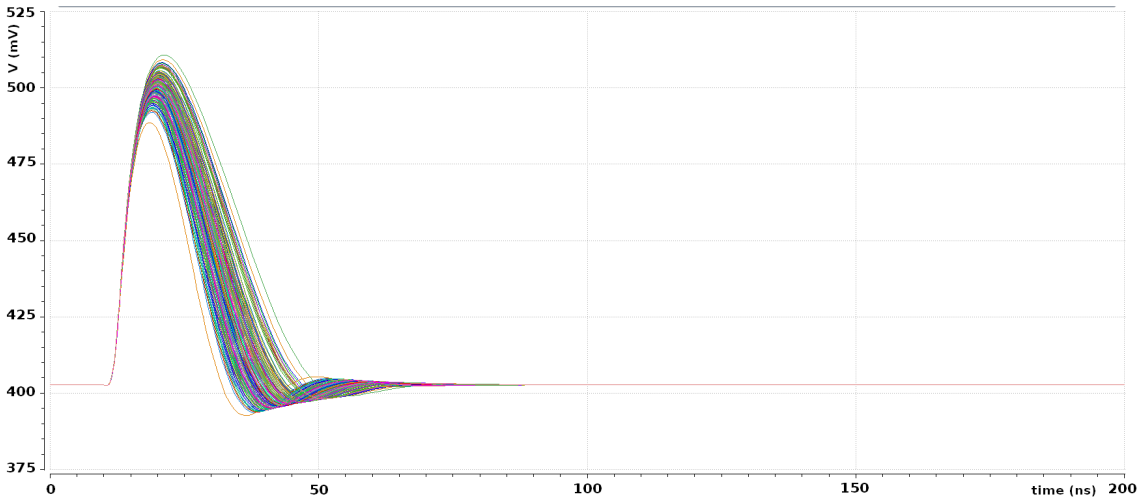


Figure 5.17: Mismatch analysis with M0 and M1 only

- **CSA and M4** This part of the circuit, instead, affects only the baseline value. In addition, it is appropriate to observe that these baseline variation are not dramatically huge, then the situation can be considered satisfying for the moment and this part of the circuit will not be modified.

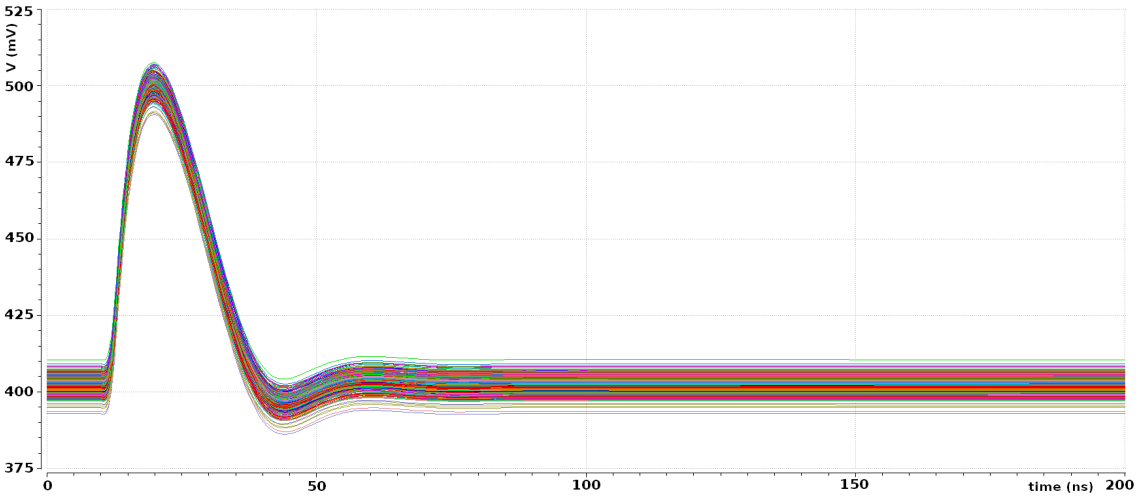


Figure 5.18: Mismatch analysis with the CSA and M4 only

All the other elements give a significant but smaller contribution. Then, since the overall area consumption has to be kept as small as possible, only the W and L of M0, M1, M2 and M3 are doubled. The new aspect ratios are listed in Table 5.3. Now the mismatch simulation of the whole preamplifier is repeated with the new transistor sizes.

(a) Transistors aspect ratio

Transistor	$W(\mu m)$	$L(\mu m)$
$M0$	0.5	2
$M1$	0.5	2
$M2$	1	1
$M3$	1	1
$M4$	1.5	0.3
$M5$	0.5	1
$M6$	0.25	1

Table 5.3: Values of the circuit parameters

As expected, the situation showed in Figure 5.19 is better than the previous one. In fact, the fluctuations of the signal time width are reduced.

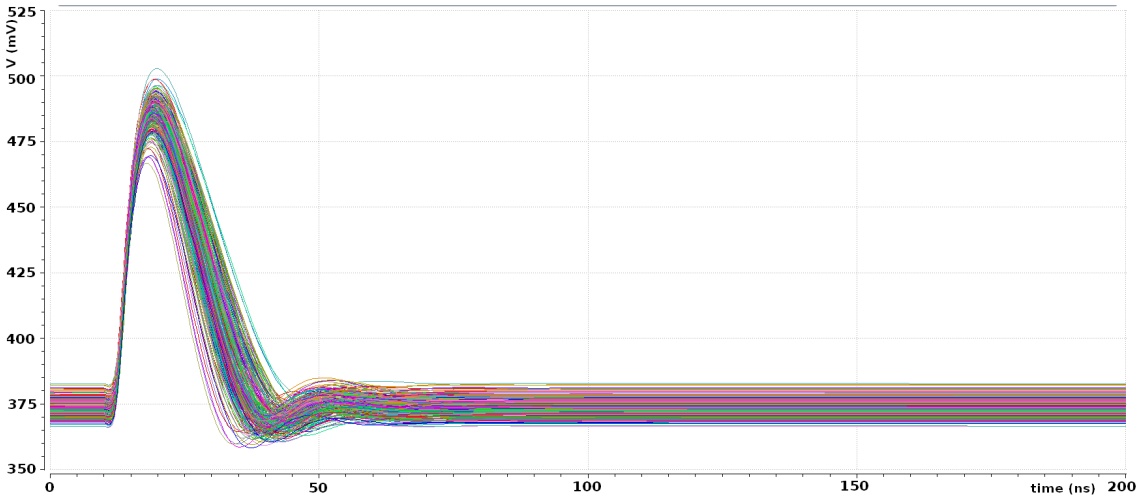


Figure 5.19: Mismatch analysis of the modified architecture

5.2.2 Second architecture

The second feedback implemetation analyzed is shown in Figure 5.20. The structure is taken by the well known Krummenacher scheme [14], although for the moment the leakage compensation part is neglected.

The main difference with respect to the first architecture is that this one is already bipolar. It is possible thanks to the differential pair formed by M2 and M3. Nevertheless, this feedback implementation requires a greater number of transistors with respect to the previous one.

The performed analyses are the same made for the previous architecture, in order to make a comparison between the two circuits. The starting transistor sizes are shown in the Table 5.4.

(a) Transistors aspect ratio

Transistor	$W(\mu m)$	$L(\mu m)$
$M0$	0.3	1
$M1$	0.3	1
$M2$	0.3	1
$M3$	0.3	1
$M4$	0.3	1
$M5$	0.3	1

(b) Current and capacitance values

I_{bias}	100nA
$C_{detector}$	100fF
$C_{feedback}$	5fF
V_{ref}	300mV
V_{pol}	360mV

Table 5.4: Values of the circuit parameters

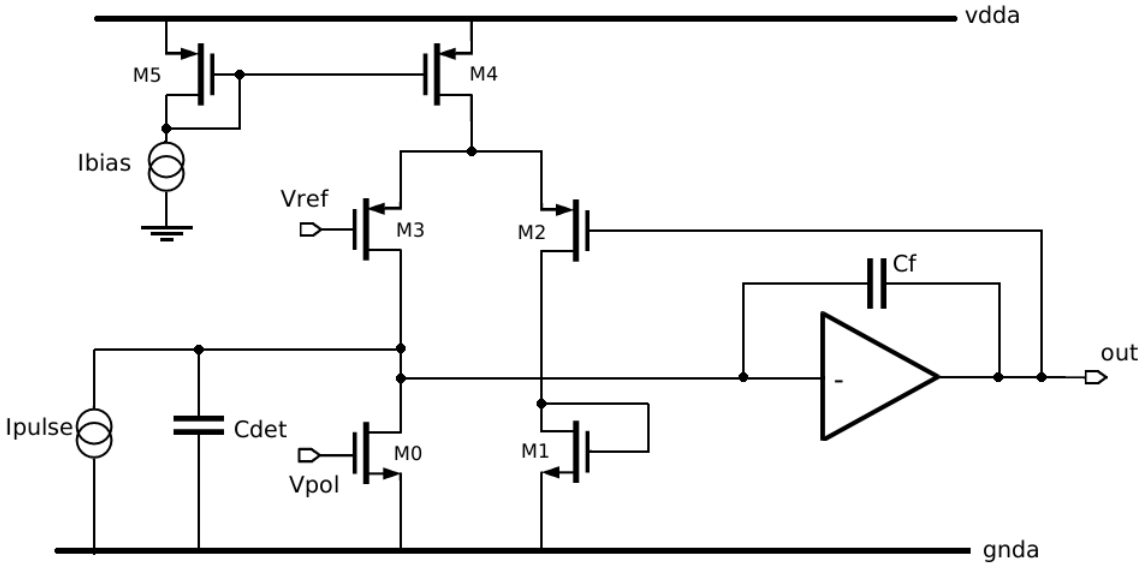


Figure 5.20: Schematic of the second architecture

It is important to underline that usually the Krummenacher scheme is used with bias currents up to $10 - 20nA$. Nevertheless, this choice leads to signals with a too long duration. In the CMS upgrade, in fact, the particle rate will be higher than in the past, requiring a faster Front-End chip. For this reasons in this case the bias current is $100nA$. In addition, as it will be better explained afterwards, the choice of a small bias current makes this circuit unipolar from the leakage compensation point of view. Obviously, it is necessary to understand if the higher bias current gives rise to new issues. In fact, it leads to an increase of the parallel noise contribution of M0. Therefore, it is necessary to check how much it affects the ENC value. Using the formula introduced in 3.2.3 and that the g_m of M0 is $42.65ns$ with IC = 0.07:

$$ENC = \sqrt{\frac{4kT\gamma g_m}{q^2}\frac{e^2}{8}T_P} = 11e^- \tag{5.7}$$

Therefore, since the noise value found in chapter 4 was around 100-110 electrons, this can be considered an acceptable contribution.

Power consumption evaluation

Also in this case an evaluation of the power consumption has been made. The total static power consumption is $2.1\mu W$. Therefore, it is almost identical to the value found for the first architecture.

Gain linearity

The gain linearity has been carried out in the same way as before. The results are shown in Figure 5.21.

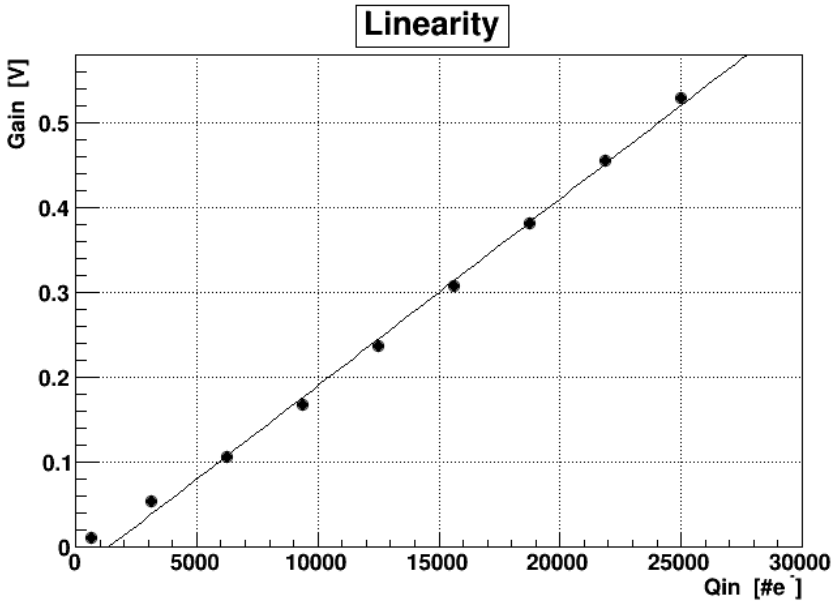


Figure 5.21: Gain vs input charge

Figure 5.21 shows that, for input charge values below 4000 electrons, there is a small deviation from the ideal trend. Apart from that, the gain linearity is well verified also in this case.

Transient noise analysis

The transient noise analysis has been performed for an input charge of $1fC$, i.e. 6250 electrons, but both positive and negative, in order to verify the bipolarity of the circuit. The two family plots are shown in Figure 5.22 and 5.23.

Since in both cases the output signal amplitude is around 125 mV and both peaking time and signal duration are very similar, it is possible to say that this

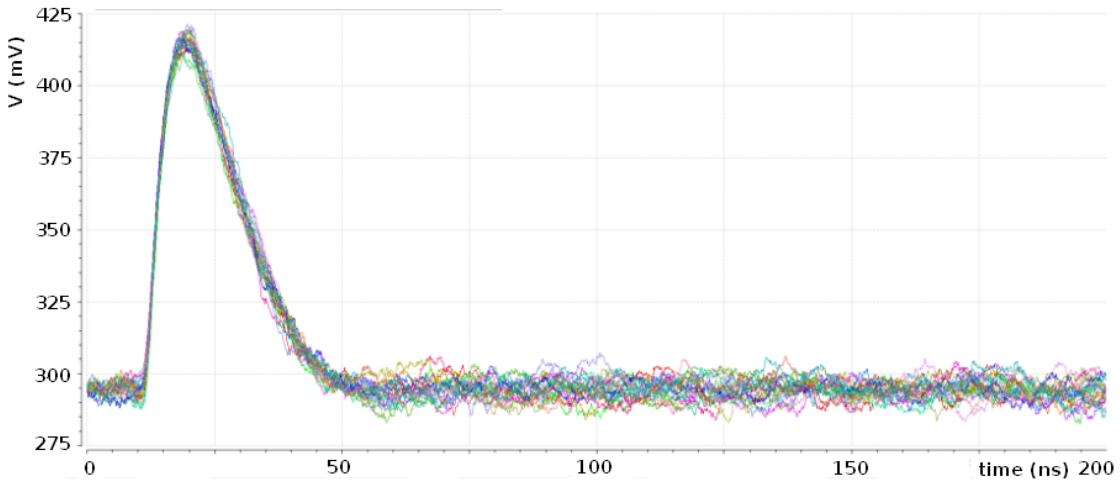


Figure 5.22: Transient analysis for a negative input signal

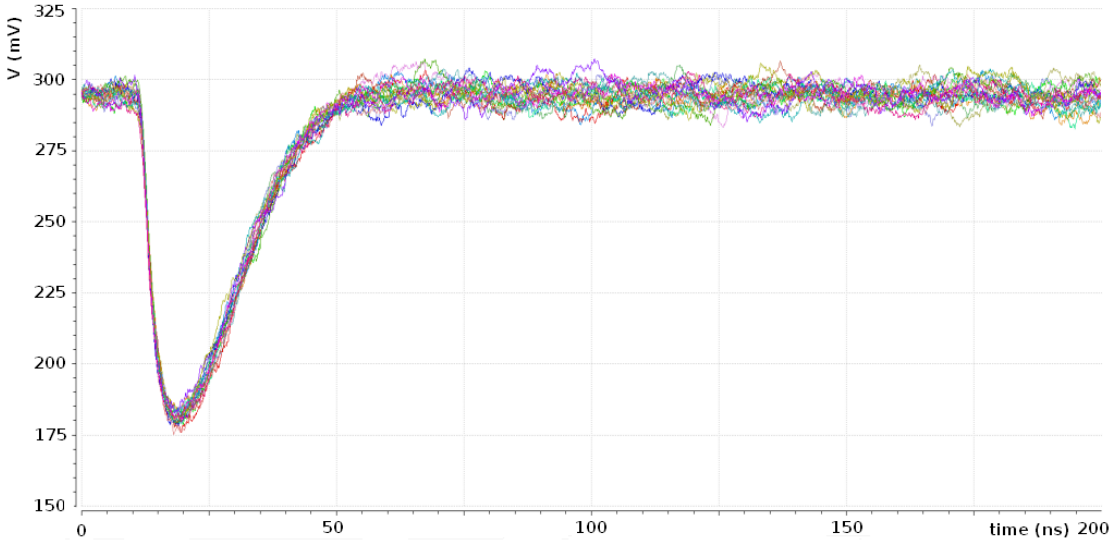


Figure 5.23: Transient analysis for a positive input signal

architecture is actually bipolar. Considering the positive signal, the histogram of the baseline values is reported in Figure 5.24.

So in this case the RMS noise is $3.1mV$. Converting the value in units of electron charge:

$$RMS[electrons] = \frac{6250 \times 3.1}{110} \simeq 180e^{-} \tag{5.8}$$

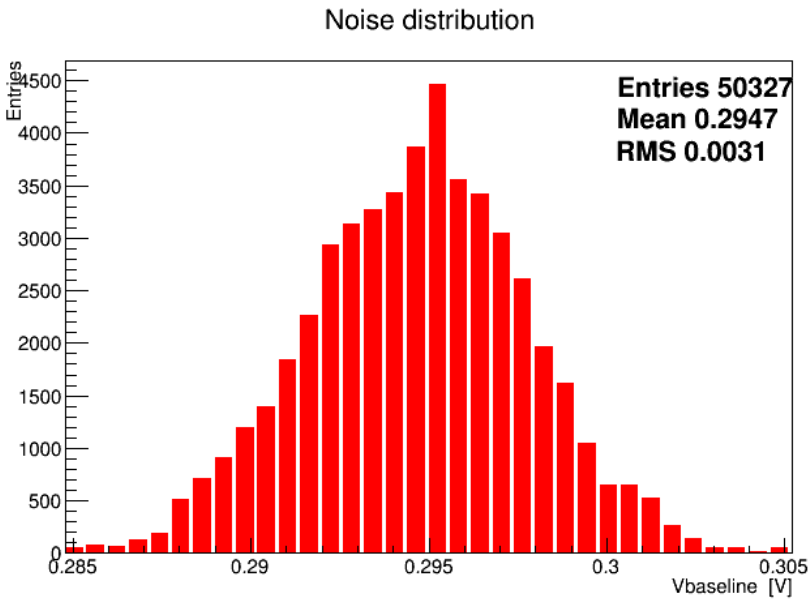


Figure 5.24: Histogram of the baseline fluctuations due to noise

As a consequence, the threshold value for this architecture is:

$$Threshold = 5 \times RMS[electrons] = 900e^{-} \tag{5.9}$$

This value is greater than the one previously found, but it is acceptable because it is below the limit value necessary for the upgrade, which is 1000 electrons.

Mismatch analysis

At this point a Monte Carlo simulation of the mismatch effects has been performed, using the transistor aspect ratios reported in the Table 5.4. The result is shown in Figure 5.25.

With this configuration there are huge fluctuations of baseline values and signal duration. In order to strongly decrease these effects, a first step is to double the W and L values of all the transistors, although it leads to a greater area occupation.

Transistor	$W(\mu m)$	$L(\mu m)$
$M0$	0.6	2
$M1$	0.6	2
$M2$	0.6	2
$M3$	0.6	2
$M4$	0.6	2
$M5$	0.6	2

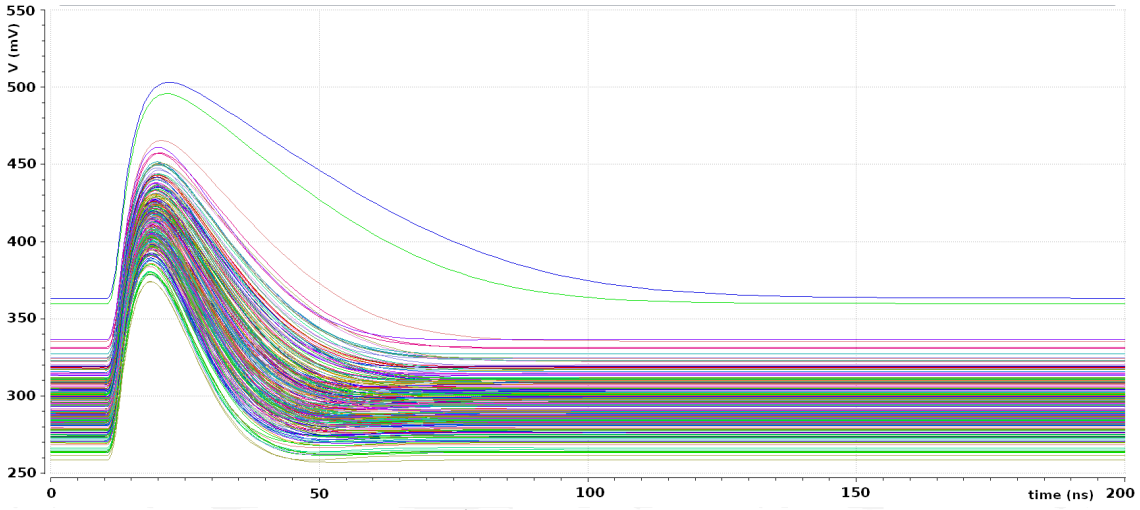


Figure 5.25: Mismatch simulation

At this point the Monte Carlo simulation has been repeated. As shown in Figure 5.26, the interval of variation of the baseline is almost halved ($50mV$ against $100mV$), and also the fluctuation of the signal duration is reduced.

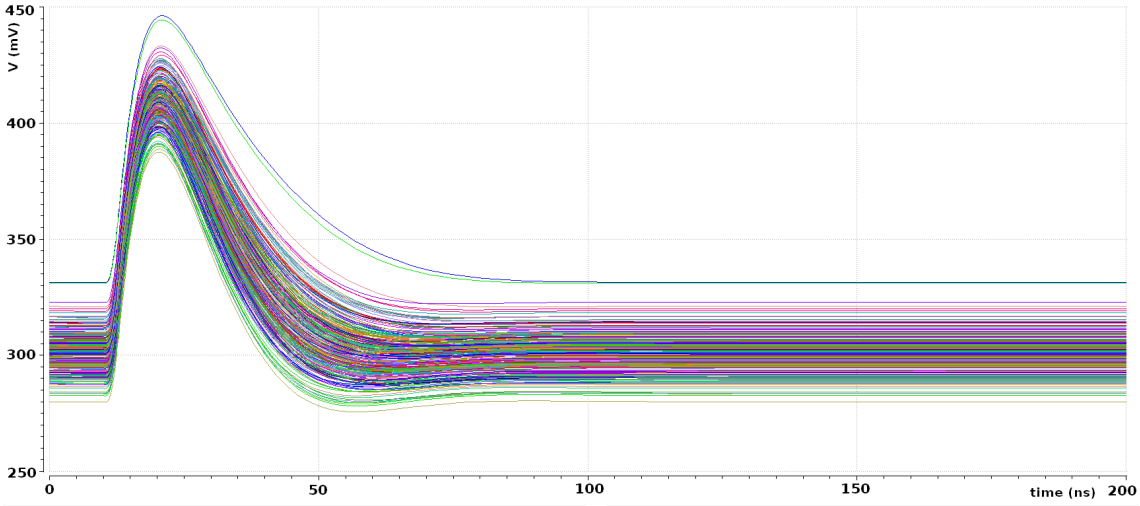


Figure 5.26: Mismatch simulation with double-sized transistors

Nevertheless, this situation is not yet satisfactory. The mismatch contributions have to be further reduced in order to avoid an excessive difference of behavior between one chip and another. The first thing to do is then to check which part of the

circuit gives the main mismatch contribution, performing Monte Carlo simulations of only some of the transistors.

- **M0 and M1** Figure 5.27 points out that these two transistors form a relevant mismatch source, especially for the baseline variation.

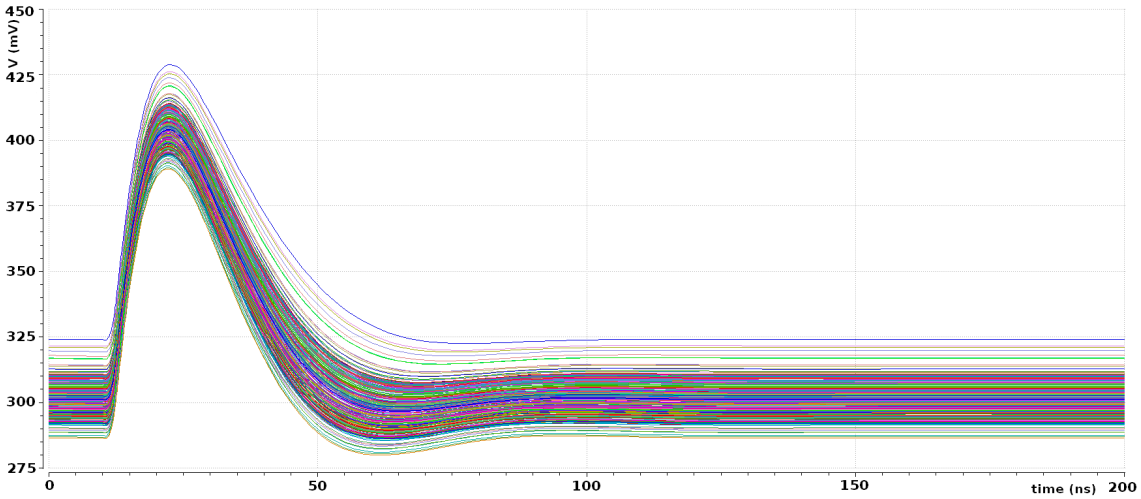


Figure 5.27: M0 and M1 mismatch simulation

- **Current mirror (M4 and M5)** Also the current mirror gives important mismatch contributions, as shown in Figure 5.28, although they are a little less important than in the previous case.

- **Differential pair (M2 and M3)** The situation of the differential pair is better: in fact, as it is possible to see in Figure 5.29, the discrepancy between one shape and another is quite small.

- **Charge Sensitive Amplifier** Figure 5.30 demonstrates instead that the Charge Sensitive Amplifier has no impact on the baseline and signal duration variation. As a consequence, it is not necessary to make further optimizations on it.

From this analysis it is clear that to improve the situation it is necessary to work on M2 and M3 and on the current mirror. One element which is important to remark is that, from the mismatch point of view, current mirrors have to operate in the strong inversion region. It is then appropriate to maintain the channel area of the transistor, but increasing the value of the channel length while reducing the channel width. This choice avoids an excessive enlargement of the area occupied by

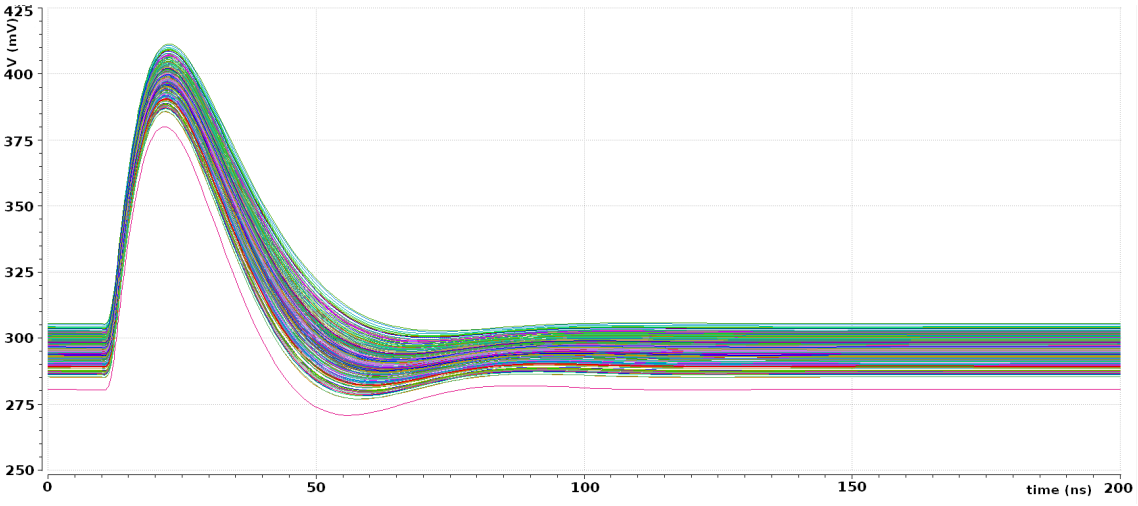


Figure 5.28: Current mirror mismatch simulation

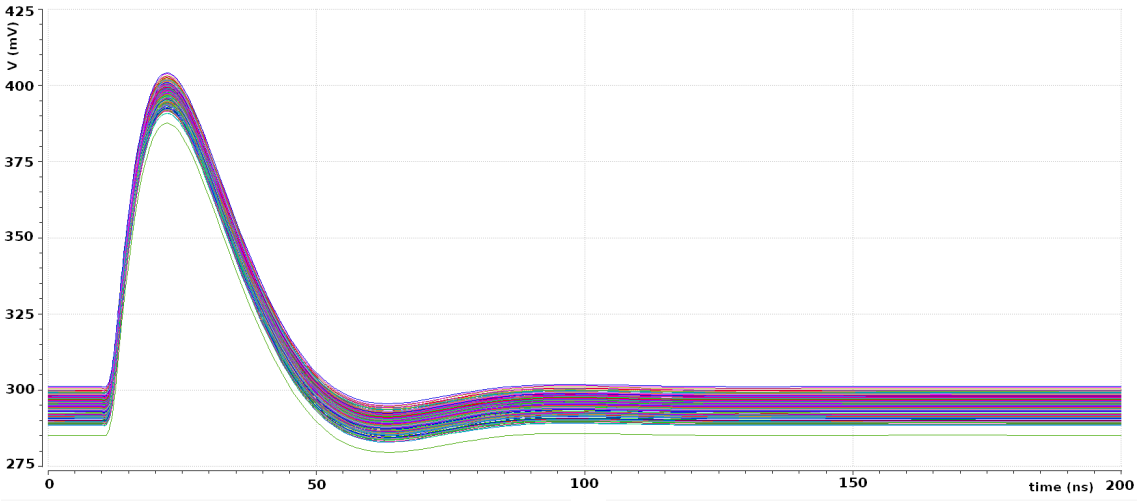


Figure 5.29: Differential pair mismatch

the circuit. In addition, recalling the expression of the inversion coefficient, which discriminates between the different operating regions

$$I_C = \frac{I_D}{2n\mu C_{ox}\frac{W}{L}V_T^2} \tag{5.10}$$

it is possible to see that decreasing the $(\frac{W}{L})$ value the inversion coefficient increases, moving towards the strong inversion region. These considerations lead to the new transistor sizes reported in Table 5.5.

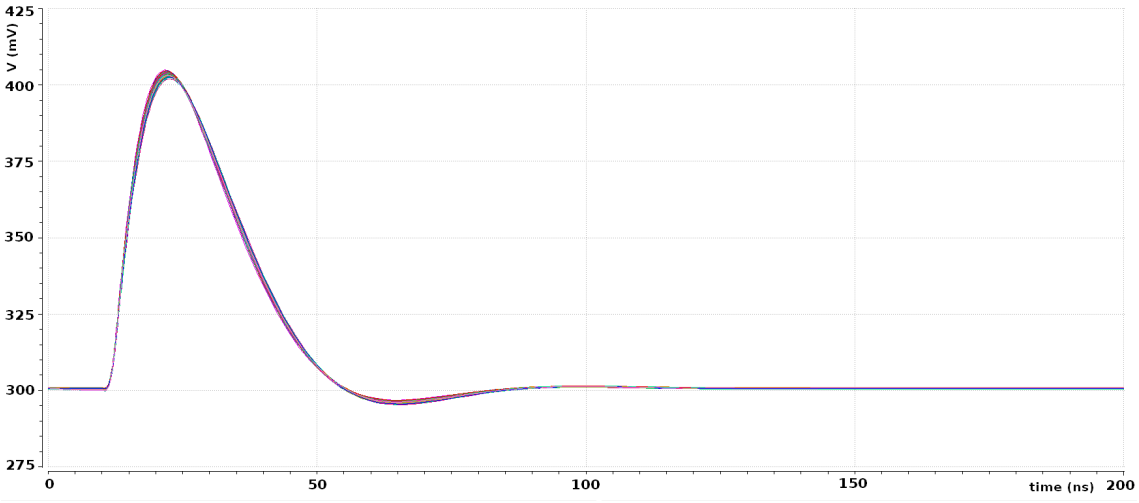


Figure 5.30: First stage mismatch

Transistor	$W(\mu m)$	$L(\mu m)$
$M0$	0.2	6
$M1$	0.2	6
$M2$	0.6	2
$M3$	0.6	2
$M4$	0.2	6
$M5$	0.2	6

Table 5.5: Aspect ratios of the new transistors

Figure 5.31 shows that, thanks to the new transistors aspect ratios listed in Table 5.5, the mismatch effects are strongly cut down with respect to the case shown in Figure 5.26. In fact, both the baseline and the signal duration variation are moreover halved.

This configuration appears to be the best possible compromise. In fact, it is not possible to further reduce the channel width of these transistors because $200nm$ is the minimum allowed width. On the other hand, it is not appropriate to further increase the overall circuit area. The readout chip, in fact, has to stay into the single pixel size. Therefore the area of the preamplifier, which is only the first stage of the analog chain, has to remain well below these limits.

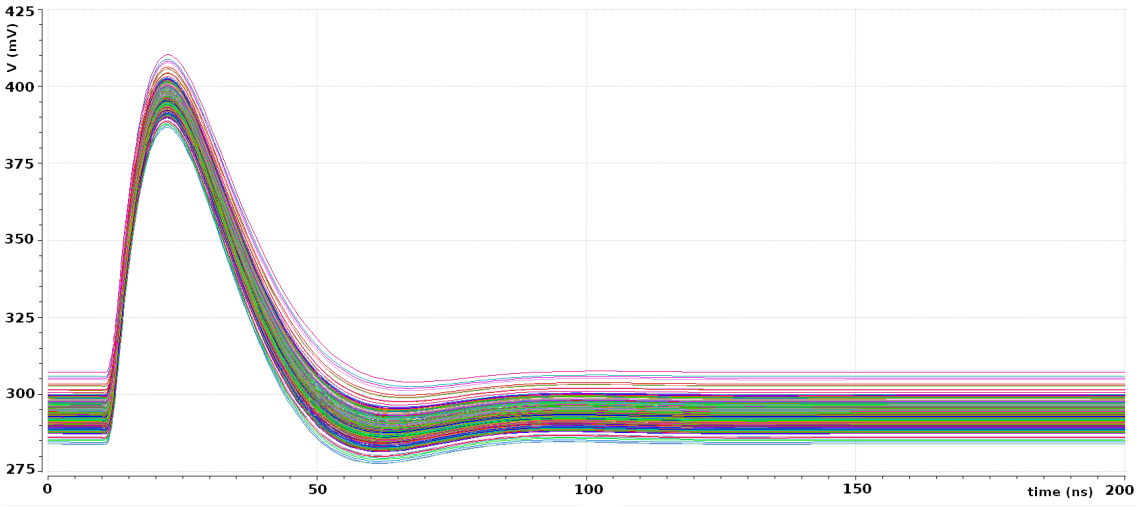


Figure 5.31: Mismatch simulation of the modified architecture

5.2.3 Summary

Putting things together, these results show that the main circuit parameters are very similar in the two cases. In fact, the power consumption and the linearity are comparable. In addition, the second architecture is intrinsically bipolar and allows quite easily the implementation of the leakage compensation part. The other scheme, instead, becomes bipolar with appropriate modifications and requires an additional block for the leakage compensation. Nevertheless, the feedback network of the latter requires a smaller number of transistors, leading to a reduction of mismatch effects.

5.3 Leakage compensation

In order to perform the leakage compensation the bipolar architecture is modified as shown in Figure 5.32.

As mentioned before, this architecture, which includes also the leakage compensation part, is the Krummenacher scheme [14]. In the choice of the transistors sizes the idea is to maintain the values determined in the previous part, as reported in the Table 5.6.

The interesting aspect of this circuit is that, with the chosen bias current value, also the leakage compensation is bipolar. In other words, it compensates both positive and negative leakage currents, provided that they remain below $50nA$. Going into details, if the leakage current is positive is not a problem, because the generator I_{leak} pushes the current into M0 so it directly returns to ground without affecting the behavior of the rest of the circuit. If instead the current is negative, the generator

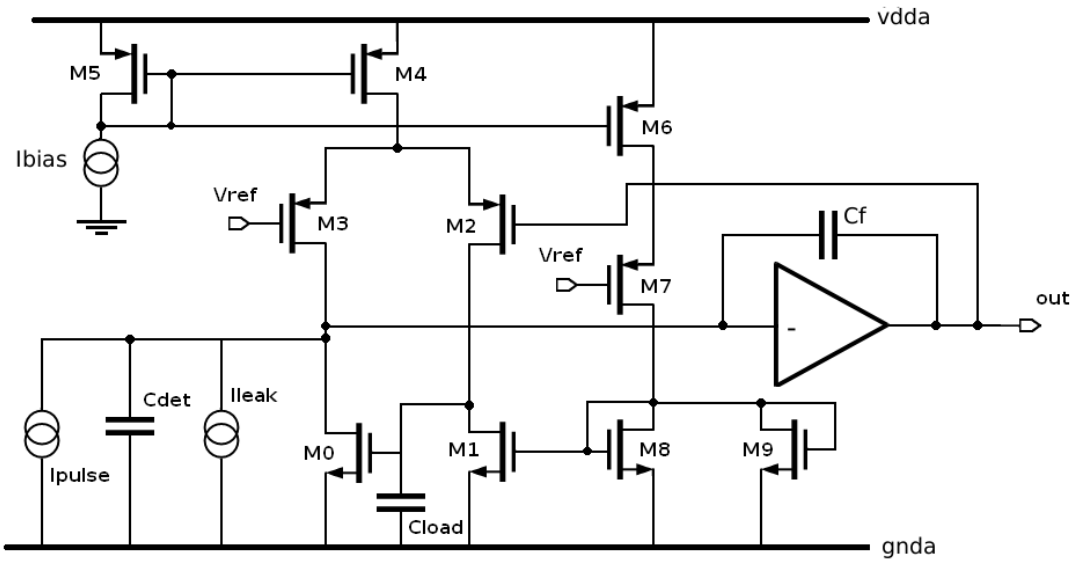


Figure 5.32: Leakage compensation architecture

(a) Transistors aspect ratio

Transistor	$W(\mu m)$	$L(\mu m)$
$M0$	0.2	6
$M1$	0.2	6
$M2$	0.6	2
$M3$	0.6	2
$M4$	0.2	6
$M5$	0.2	6
$M6$	0.2	6
$M7$	0.2	6
$M8$	0.2	6
$M9$	0.2	6

(b) Current and capacitance values

I_{bias}	100nA
$C_{detector}$	100fF
$C_{feedback}$	5fF
C_{load}	200 fF
V_{ref}	300mV

Table 5.6: Leakage compensated circuit parameters

starts to pull the current from the branch of M0. Since in each of the two branches of the differential pair there is a current of $50nA$, when the leakage current rises over this value the differential pair starts to be very unbalanced. In addition, when the leakage current is around $100nA$, the leakage generator takes all the bias current, and the circuit does not work anymore. Therefore, this architecture is optimized to

compensate leakage currents up to $50nA$. This is an acceptable limit, because at the beginning of the sensor operation the leakage current is very small, below $1nA$, but it rises with the radiation damage. Nevertheless, it is not expected to significantly overcome this value.

Signal form

The first thing to do is to check how the signal form changes applying a leakage current. In this paragraph some particularly significant cases are reported. The first analysis, reported in Figure 5.33, is made neglecting the leakage current, in order to verify if the circuit works well in ideal conditions.

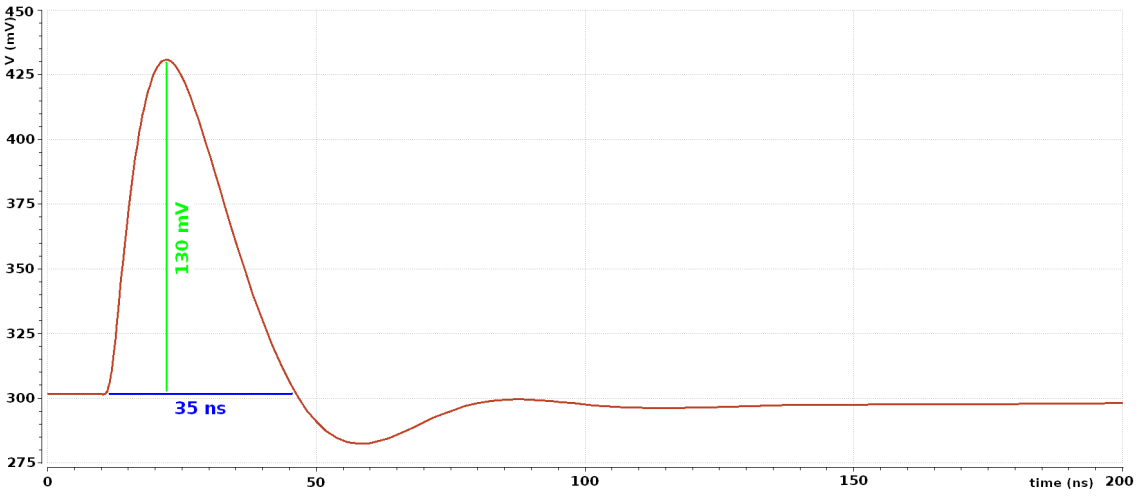


Figure 5.33: Output signal (leakage current = $0nA$)

It is then interesting to verify the situation when the leakage current is present. For example, the case with $10nA$ is reported in Figure 5.34.

Finally, it is appropriate to control if the output signal remains good also with large leakage currents, for example $90nA$, a value which however will be hardly reached, and, in case, only in the last part of the sensor life. The result is shown in Figure 5.35.

From this three plots it is clear that positive leakage currents up to around $100nA$ have a very little influence on the output signal, in fact there is no significant variation of amplitude, duration and undershoot-overshoot of the signal.
In addition, the same analysis has been performed with negative current values, in order to check if the expected bipolarity is verified.

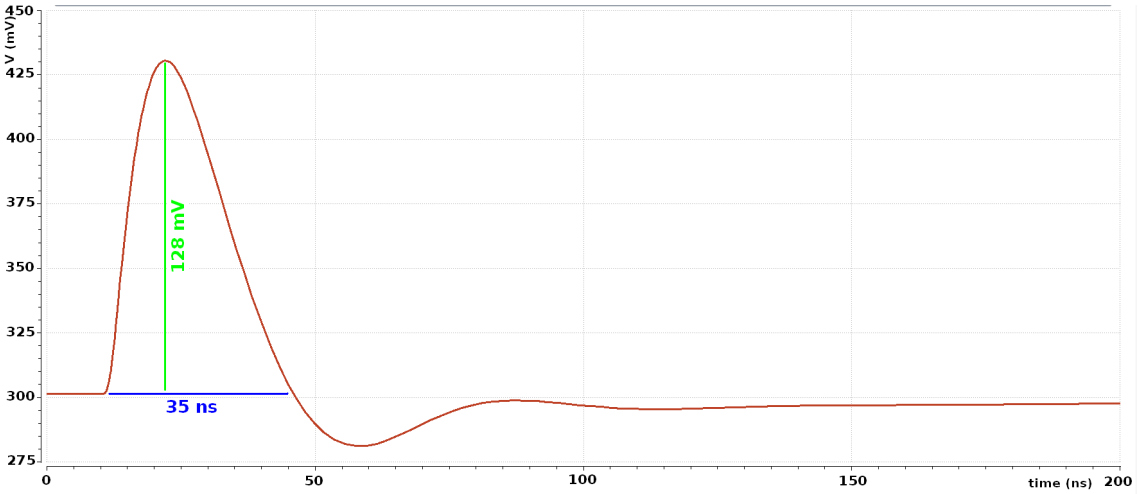


Figure 5.34: Output signal (leakage current = $10nA$)

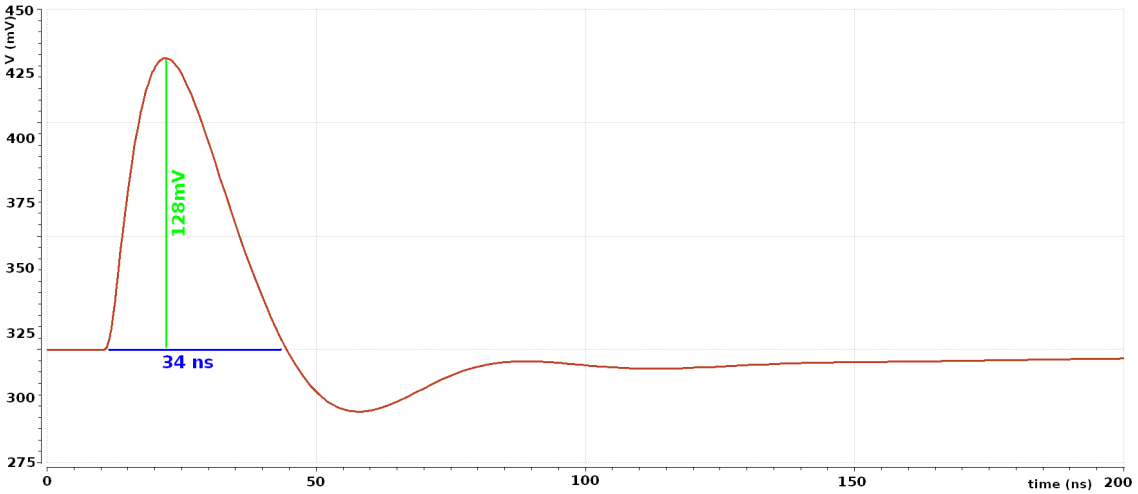


Figure 5.35: Output signal (leakage current = $90nA$)

Figure 5.37 confirms that with negative currents, the leakage compensation is not as good as with positive currents. In fact, with $I_{leak} = -90nA$ the signal duration is hugely increased. It is then useful to repeat the analysis with some different leakage current values in order to better understand the limit beyond which the compensation of negative currents is not good. This analysis is presented in Figure 5.38.

It is clear that with a leakage current up to $I_{leak} = -50nA$ the output signals are

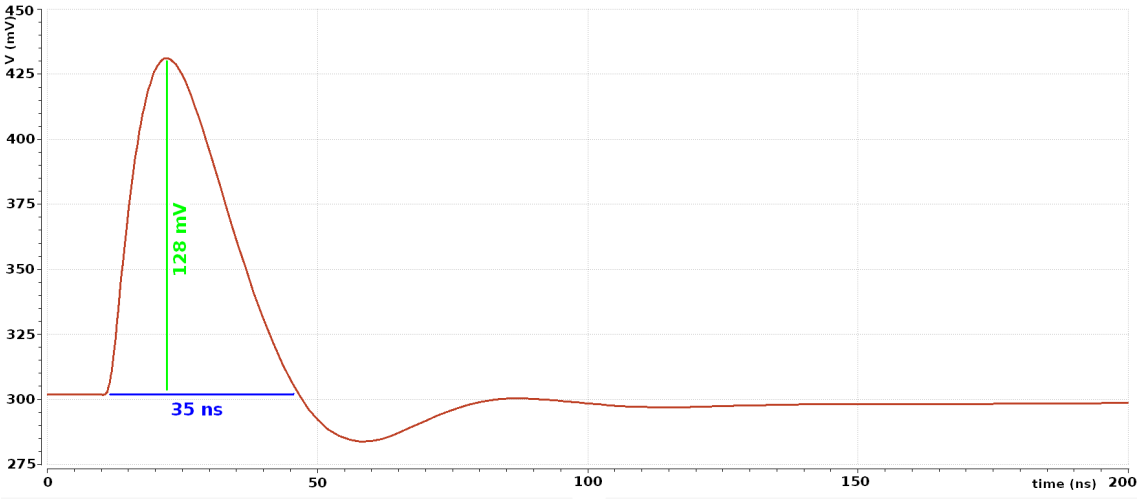


Figure 5.36: Output signal (leakage current = $-10nA$)

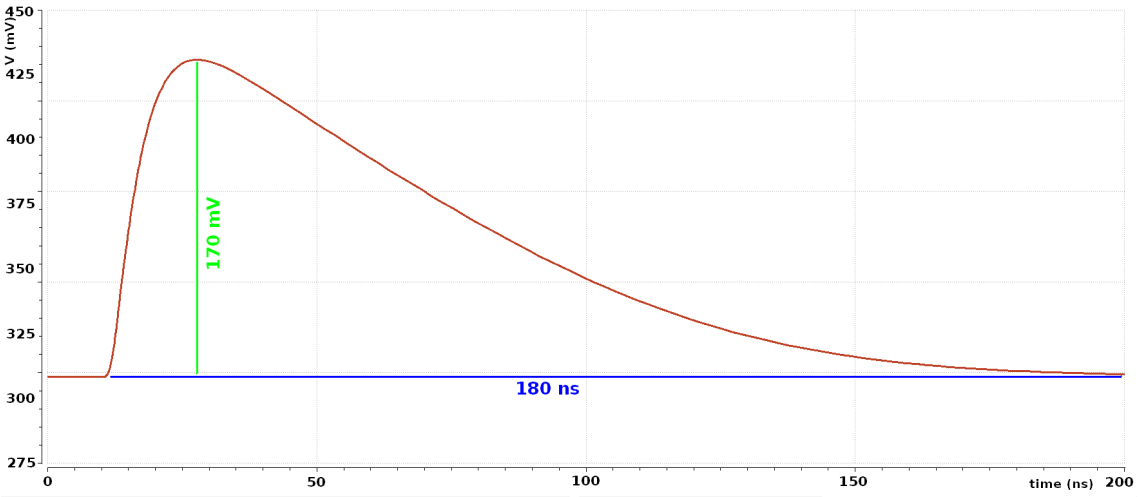


Figure 5.37: Output signal (leakage current = $-90nA$)

very similar, therefore the compensation is good. Beyond these values the compensation progressively becomes less effective. However, as explained before, it has to be taken into account that leakage currents beyond $50nA$ will be hardly reached. Then these results can be considered quite satisfying and consistent with the expectations.

At this point it is necessary to check some other aspects of the circuit in order to verify its performances, assuming that the leakage current has a value which is well compensated, so that it does not significantly affect the output signal.

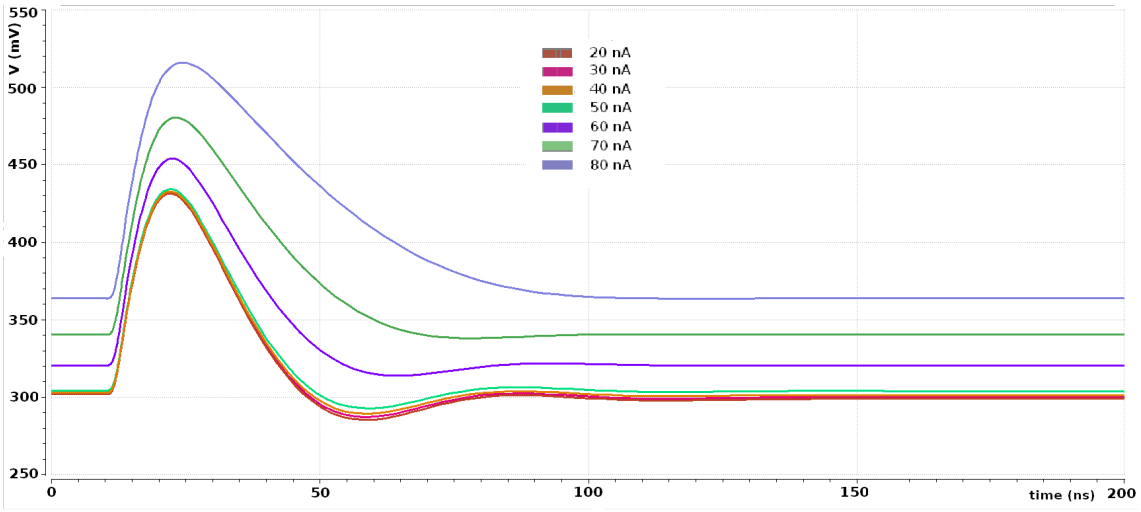


Figure 5.38: Output signals for different negative leakage currents

Power consumption estimation

This analysis has been performed with the same technique used in the previous cases. The static power consumption is $2.3\mu W$, which is a little higher that the value found without leakage compensation (which was $2.1\mu W$). This difference is explained by the fact that in the final architecture a branch in which flows a current of $100nA$ has been added.

Mismatch analysis

Since this architecture contains more transistors than the version without leakage compensation, it is appropriate to verify if the influence of mismatch effects on the output signal is changed. The results of the analysis are shown in Figure 5.39.

As it could be expected, comparing the situation with the case without leakage compensation there is a small increase of both baseline and signal duration variation. Nevertheless, since the variations are not very relevant, this result can be considered a good compromise, since a further optimization would lead to an excessive increase of the area consumption.

Time over threshold

The study of the Time over Threshold (ToT) is useful to underline some other crucial aspects of this architecture. Also in this case the first step is the transient noise analysis, useful not only to fix the threshold value, but also to check the noise contributions of this architecture.

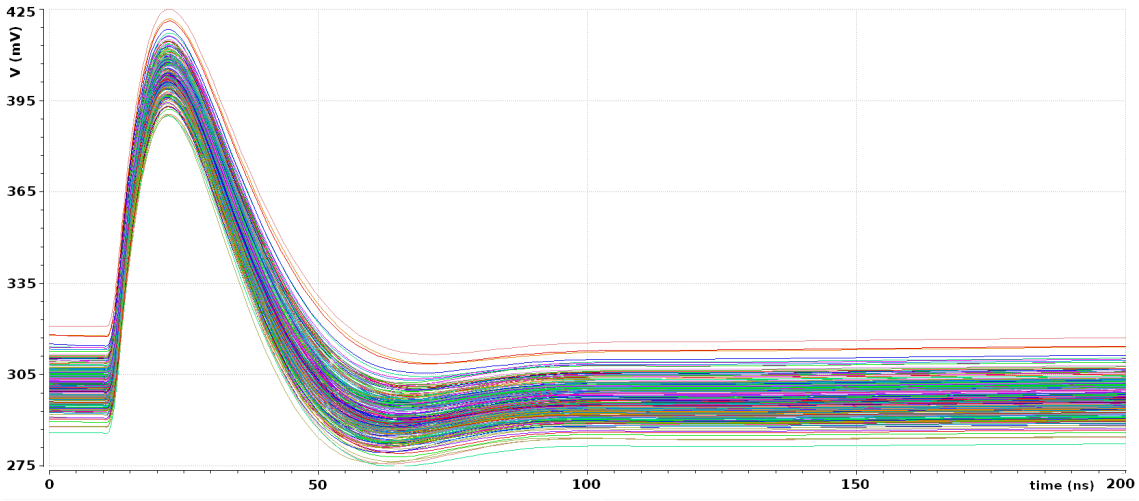


Figure 5.39: Mismatch analysis of the leakage compensated circuit

Consequently, the histogram of the baseline values has been carried-out and the

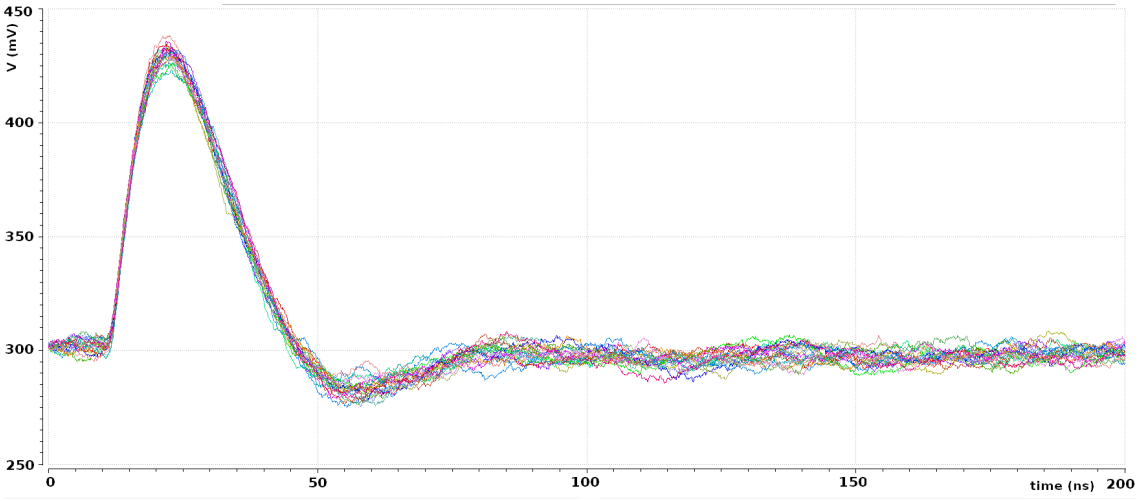


Figure 5.40: Transient noise analysis

obtained distribution is shown in Figure 5.41. This time the gain is around $130mV$, so the RMS expressed in electrons is:

$$RMS[electrons] = \frac{6250 \times 3.7}{130} \simeq 180e^{-} \tag{5.11}$$

Therefore, this result confirms that the noise value is substantially the same found without leakage compensation.

It is foreseen that the ToT increases with the input charge. In fact, both the

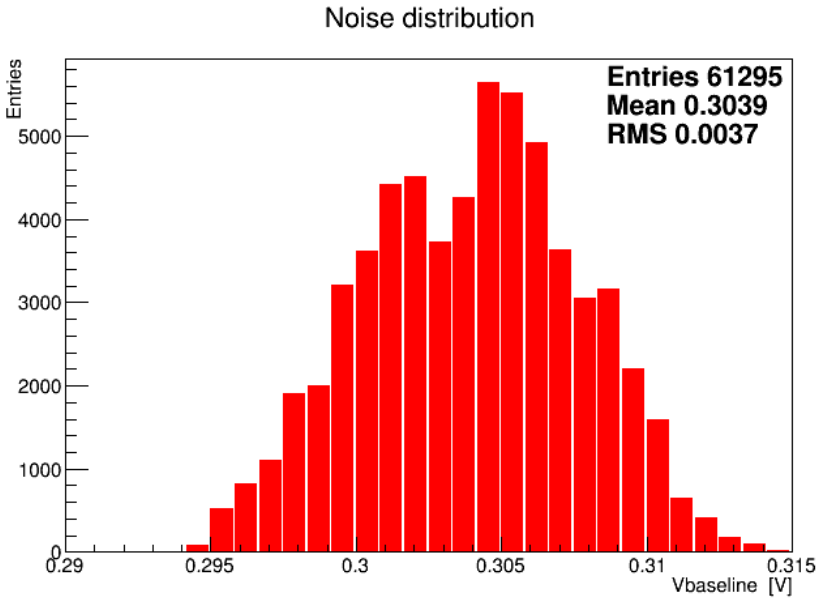


Figure 5.41: Histogram of the baseline values of the leakage compensated architecture

amplitude and the duration of the output signal are expected to linearly increase with the input charge. In order to check if these situation is verified a simulation of the ToT with different values of the input charge has been performed. The results, reported in Figure 5.42, indicate that the linear growth of the ToT is well verified in a large spectrum of input charge values.

5.3.1 Summary

In conclusion, it is possible to say that this leakage compensated architecture works as expected, in fact is allow the compensation of both positive and negative leakage currents up to $50nA$, and positive currents even higher, also if they will hardly reached. Another positive aspect of this architecture is that the addition of other transistors due to the leakage compensation part does not significantly affect the circuit performance in terms of mismatch and power consumption.

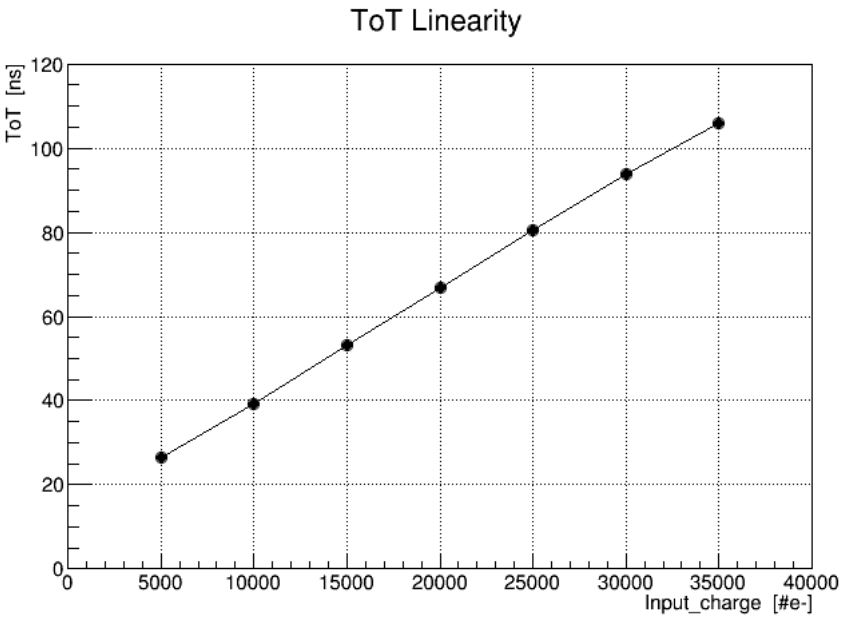


Figure 5.42: ToT vs input charge

5.4 Prospects for time-variant integrators

Sometimes, the readout of radiation sensors is performed with schemes different from the time-invariant filters described until now. In other words, also systems in which the transfer function changes during the circuit operation, called "Time-variant filters" , are used. This kind of architecture was very commonly used in many applications in the past, but not for the readout of radiation sensors. In fact, since the transistor sizes were significantly higher, these kind of architecture gave rises to some issues, such as charge injection. The circuit operation, in fact, is regulated by CMOS switches. When they are closed some charges are collected in the inversion layer. As a consequence, when the switches they are opened a part of this charge goes to the input, inducing an error on the voltage value. Nowadays, however, the development of the deep submicron technologies allows to use very small transistors, leading to a strong reduction of the charge injection. Therefore, the use of time-variant filters can be considered again.
The circuit of Figure 5.43 is an example of a time-variant charge sensitive amplifier.

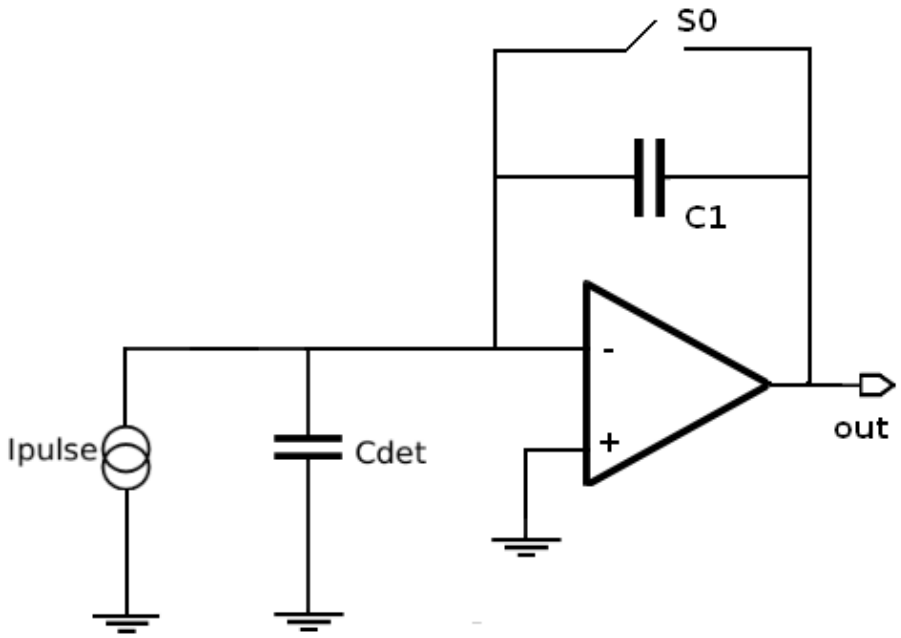


Figure 5.43: Schematic of a time-variant filter

The first thing to check is if without an input signal the baseline value remains constant. Therefore, it is necessary to realize a transient time simulation on a quite long time period, for example $1ms$. The circuit is resetted at the beginning of the simulation in order to fix the initial baseline value. The results is shown in Figure 5.44.

There is clearly a long time constant discharge. Therefore it is interesting to better understand what happens over a shorter time, in order to fix the minimum time step between one reset and another. Over a period of 100 μs, as shown in Figure 5.45, the baseline lowers of around $2mV$. Therefore, if one resets the stage every $100\mu s$ the baseline remains almost steady.

Although the transistor aspect ratio is strongly reduced with respect to the past, it is necessary to check if charge injection is still relevant also with the 65nm technology. This phenomenon is expected to be proportional to the rise time of the control signal, so the same analysis is performed with a control signal with rise time of 1ps and 1ns. In both cases the switch is closed at t=10ns and is opened at t=40ns. In addition, an input signal starts at t=50ns in order to verify the duration of the output signal. The case with rise time of 1 ps is shown in Figure 5.46.

The case with an input signal rise time of 1 ns is shown in Figure 5.47.

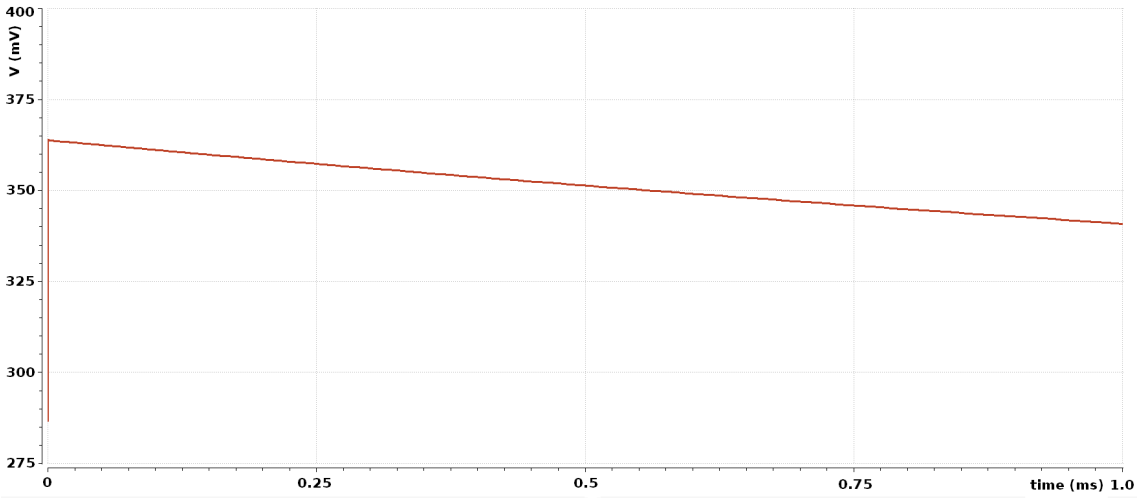


Figure 5.44: Output signal shape (time = 1 ms)

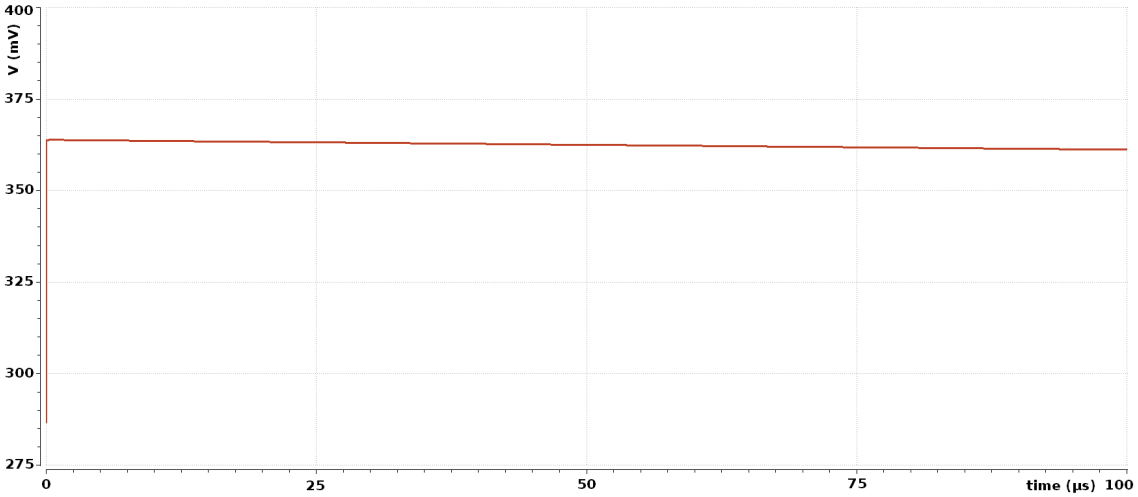


Figure 5.45: Output signal shape (time = 100 μs)

Therefore, as expected, if the rise time is higher the charge injection at the switch opening is lower. These two figures also show that the rise time of the output signal is around 20 ns. Therefore, it is comparable with the value of the peaking time of time-invariant filters. Therefore also from the signal duration point of view the choice of a time-variant filter can be considered.

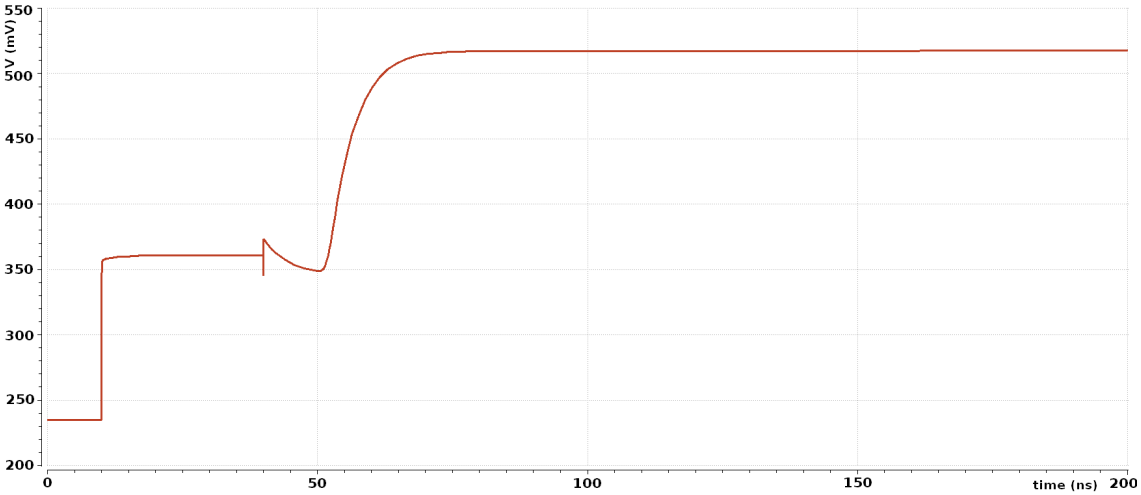


Figure 5.46: Output signal shape (input signal rise time = 1ps)

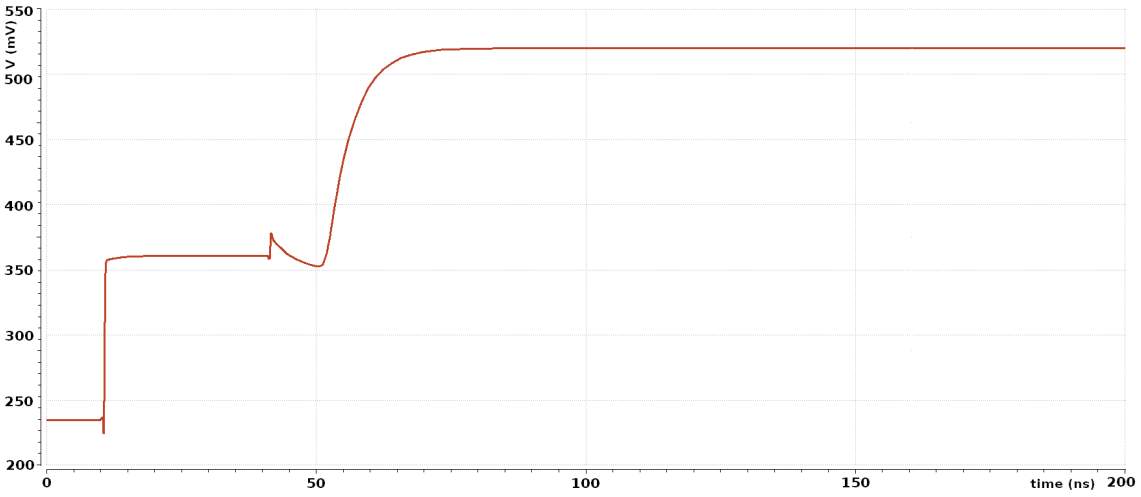


Figure 5.47: Output signal shape (input signal rise time = 1ns)

Chapter 6

Conclusions

Nowadays silicon detectors systems have to cope with the huge amount of signals coming from the sensor, avoiding the pile-up or the loss of the signals. Therefore, they need a fast readout electronics. Usually the clock of the system works at tens of MHz. This chip is realzed with a modern CMOS technology.
Furthermore, the upgrade of the silicon pixel detector of the CMS experiment leads to more severe requirements for the readout chip in terms of area consumption, radiation tolerance, power budget and noise performances. This situation led to the choice of a 65nm CMOS technology. For this reason, a new design of the analog part of the chip has to be carried-out.
Since the new chip will be realized using the 65nm CMOS technology while the old one was built with the 250nm CMOS technology, a comparison between them has been performed in order to better understand the different behavior of a single transistor in the two cases. The results show that the main parameters, such as threshold voltage and intrinsic gain, have similar values, but the advantage of the 65nm is the increase of the circuit density and, contemporarily, the drop of power consumption in the digital part. Therefore, since the 65nm maintains and, in some cases, improves the performance of the previous technologies, it is suitable for an analog and digital Front-End implementation.
Subsequently, since the signal-to-noise ratio is a crucial parameter in Front-End design, the noise optimization of a simple preamplifier stage in 65nm CMOS technology has been studied. With the chosen architecture, the main noise contribution is due to the input transistor. As a consequence, the main purpose of the analysis was to find out the aspect ratio of this device that minimized the noise contribution. In addition, the sensor capacitance and the current flowing in the stage affect the noise performance, requiring a trade-off between these quantities.
Then, in order to deal with more realistic architectures, two different feedback network implementations have been studied. Each of them has advantages and disadvantages. The first one requires only two transistors in the feedback implementation

and is better from the mismatch point of view, but it requires an additional circuit for leakage compensation. Furthermore, particular caution is needed to make it compatible with sensors of either polarity. The second topology, instead, already fulfills these requirements but it has to be used with current values higher than usual in order to be suitable for CMS, resulting in a small increment of the noise. Although the feedback network requires a higher number of transistors, this architecture allows also a leakage current compensation up to 50 nA.
Finally, also a simulation of a time-variant Front-End has been performed, showing that thanks to deep submicron technologies this kind of architecture becomes again a suitable choice.
This work shows that in principle all the three architectures can be used, but additional studies have to be made in order to discriminate between them:

- The layout of these circuits has to be studied, in order to verify which is the most compact from the area occupation point of view.
- The time-variant integrator has to be analyzed more in detail.
- It is necessary to connect the CSA to the comparator in order to build a complete analog Front-End chain

Acknowledgements

Desidero ringraziare le tante persone che mi hanno aiutato durante questo anno di lavoro per la tesi. Innanzitutto un ringraziamento particolare va al mio relatore, Dott. Angelo Rivetti, per le tante spiegazioni e suggerimenti necessari alla realizzazione di questo lavoro e per il supporto che mi ha dato in tutta la mia attività. Un sentito grazie va al Dott. Natale Demaria, che mi ha particolarmente aiutato a capire il progetto degli upgrade dei rivelatori a pixel di CMS e anche nella discussione dei risultati ottenuti in questo lavoro. Un ringraziamento speciale va inoltre a Luca Pacher, che mi ha insegnato a utilizzare in maniera appropriata il simulatore circuitale, ma che mi ha anche aiutato molto nella realizzazione e analisi delle simulazioni. Desidero inoltre ringraziare tutte le persone del laboratorio VLSI e del gruppo CMS di Torino, sempre disponibili a fornire chiarimenti quando ce n'è stato bisogno e più in generale tutti coloro che mi hanno supportato in questo periodo.

Bibliography

[1] B. Razavi, *Design of analog CMOS Integrated Circuits*. McGraw Hill, 2000.

[2] L. Rossi, P. Fischer, T. Rohe, and N. Wermes, *Pixel Detectors, from Fundamentals to Applications*. Springer, 2006.

[3] M.Borri, "Analisi dei dati di un test di fascio di rivelatori al silicio 3D FBK-irst double-side double type column," Master's thesis, University of Turin, 2009.

[4] cms.web.cern.ch.

[5] Dominguez A., "CMS technical design report for the pixel detector upgrade," tech. rep., CERN, 2012.

[6] CMS collaboration, "CMS technical design report for the pixel detector upgrade," tech. rep., CERN, 2012.

[7] CMS collaboration, "Technical proposal for the upgrade of the CMS detector through 2020," tech. rep., CERN, 2011.

[8] D. Foty, D. Binkley, and M. Bucher, "Measurements and Modeling of MOSFET Inversion Level Over a Wide Range As a Basis for Analog Design," *ECCTD'01*, 2001.

[9] F. Silveira, D.Flandre, and P. G. A. Jespers, "A g_m/i_D based methodology for the design of CMOS analog circuits and its application to the synthesis of a silicon-on-insulator micropower ota," *IEEE Journal of solid-state physics*, vol. 31, no. 9, 1996.

[10] B. Razavi, *Fundamentals of microelectronics*. Wiley, 2006.

[11] J. Hurst and R. Gray, *Analysis and design of analog integrated circuits*. Wiley, 2001.

[12] D. Johns and K. Martin, *Analog Integrated Circuit Design*. Wiley, 1996.

[13] F. Krummenacher, C.C. Enz, and E.A. Vittoz, "An analytical MOS transistor model valid in all regions of operations and dedicated to low-voltage and low-current applications," *Analog Integrated Circuits and Signal Processing*, 1995.

[14] F. Krummenacher, "Pixel detectors with local intelligence: an IC designer point of view," *Nuclear Instruments and Methods in Physics Research*, 1991.

[15] G. Lutz, *Semiconductor radiation detectors*. Springer, 1999.

[16] C.-C. Liu, S.-J. Chang, G.-Y. Huang, and Y.-Z. Lin, "A 10-bit 50-MS/s SAR ADC With a Monotonic Capacitor Switching Procedure," *IEEE Journal of solid-state circuits*, vol. 45, no. 4, 2010.

List of tables

List of figures