Università degli Studi di Torino

Dipartimento di Fisica
Corso di Laurea Magistrale in Fisica delle Tecnologie Avanzate

Analog Front-end Design in Deep Sub-micron CMOS Technology for Timing application in Pixel Detectors

Candidato:
Lorenzo Piccolo

Relatore:
Angelo Rivetti

Controrelatore:
Michela Greco

a.a. 2016-2017

Contents

Abstract **iv**

1 4D Tracking Detector for HL-HEP Experiments **1**
1.1 HL-LHC Upgrade . . . 1
1.2 The TIMESPOT Project . . . 4
1.2.1 TIMESPOT Radiation Sensors . . . 7
1.3 Electronics for Timing . . . 12
1.3.1 Timing Metrics . . . 12
1.3.2 Overview on Timing Architectures . . . 15

2 Deep Sub-Micron *CMOS* Electronics **23**
2.1 *MOSFET* Properties . . . 23
2.1.1 Electrical Characteristics and Working Regions . . . 25
2.1.2 Small Signal Parameters . . . 32
2.1.3 MOSFET Noise . . . 35
2.1.4 Mismatch in *MOSFETs* . . . 37
2.2 *CMOS* Fabrication Process and Design . . . 39
2.2.1 *CMOS* technologies . . . 40
2.2.2 Manufacturing Process . . . 41
2.2.3 Design Cycle . . . 43
2.3 Scaling to the 28nm Technology node . . . 45
2.3.1 Scaling . . . 45
2.3.2 High-k dielectric with Metal Gate . . . 47
2.3.3 Channel Strain . . . 49
2.3.4 Resolution Enhancement Techniques (RET) . . . 49
2.3.5 Impact of Scaling on Analog Circuits . . . 51
2.4 Basic *CMOS* Analog structures . . . 52
2.4.1 Buffers and Circuit Biasing . . . 53

2.4.2 Single Ended Amplification Stages 54
2.4.3 Differential Amplification Stages 58

3 Front-End Architecture 61
3.1 Overview 62
3.2 Charge Sensitive Amplifier 63
3.2.1 Core Amplifier 69
3.2.2 Krummenacher Filter 72
3.3 Leading Edge Discriminator 75
3.3.1 Amplification Stages 75
3.3.2 Offset Correction 78

4 Simulation Results 83
4.1 Comparative Technology Tests 84
4.2 CSA Tests 87
4.2.1 Core Amplifier Transfer Function 87
4.2.2 CSA Signal Characteristics 89
4.3 Leading Edge discriminator Tests 91
4.4 Timing Tests 94
4.4.1 Noise Contribution 95
4.4.2 Signal Variation Contribution 96
4.4.3 Mismatch Contributions 97
4.4.4 Time Walk 99
4.4.5 Process Variations Contributions 103

Conclusions 106

Bibliography 109

High Luminosities (HL) upgrades planned at the Large Hadron Collider (LHC) will pose strict requirements for High Energy Physics (HEP) experiments. Future vertex detector systems need to combine pixel dimensions of 100μm$\times100\mu$m or smaller with time resolution below $100ps$ in order to obtain satisfying tracking performance and prevent event pile-up. Moreover these detectors will have to sustain radiation doses of $10^{17} MeV/cm^2 n_{eq}$ in magnitude.
In this work an analog CMOS front-end architecture for pixel detectors which satisfies these requirements is proposed. This front-end will be part of TIMESPOT project, a national initiative by INFN. The project aims to develop an integrated solution for a tracking detector that meets HL HEP requirements; therefore sensors, front-end ASIC and back end electronics will be developed as part of it. As a consequence of this, the studied analog front-end design must take into account the need to be interfaced with other system's blocks.
To achieve this, the work was done in coordination with other TIMESPOT working groups. For the scope of this thesis, such coordinate effort was done mainly with sensors developing groups.
Pixel detectors electronics pose many design challenges: low power consumption, high signal-to-noise-ratio, radiation hardness and small chip area. Furthermore the timing requirement of 100ps resolution demands robust design against process variations.
The target pixel pitch was chosen to be $55\mu m \times 55\mu m$ so, to integrate all required features, a 28nm technology node was chosen for the front-end electronics. In order to efficiently design the front-end in this advanced deep sub-micron technology, characterization studies where done beforehand.
The work was carried out using industry standard tools and methodologies for design, simulation and layout of CMOS integrated circuits. Technologies libraries provided by manufacturer were employed in circuits simulation and process variation impact evaluation.
The thesis initially introduces HL-LHC challenges, the TIMESPOT project and its sensors in chapter 1, electronics techniques and architectures for timing are also illustrated. Chapter 2 illustrates the most important characteristics of these technologies by the means of electronics mathematical models and technologies details

with a particular interest in deep sub-micron CMOS technologies. Chapter 3 illustrates the designed architecture and discuss the various optimizations adopted in this work. In the end, in chapter 4, both results on a preliminary design in 65nm and the final 28nm version are presented and compared. In these analysis particular attention is paid to timing performance.

Chapter 1

4D Tracking Detector for HL-HEP Experiments

The studied front-end architecture with timing measure is part of the *TIMESPOT* project, its goal is to produce a 4D tracking system for *High Luminosity* (*HL*) accelerator experiments. In this chapter firstly, in section 1.1, motivations and requirements of *HL* experiments are briefly illustrated, with a particular focus on *HL-LHC* expansion. Then the *TIMESPOT* project is presented in 1.2, explaining its organization, systemic approach and the role of the front-end electronics in it. In 1.2.1 the work in progress realization of the two 3D sensor considered in *TIMESPOT* are presented due to the strict connection between their design and the front-end one. Finally in 1.3 timing metrics and architectures are presented, their feasibility in respect of the desired front-end is also discussed.

1.1 HL-LHC Upgrade

CERN (Conseil Européen pour la Recherche Nucléaire) is the world leader organization in *HEP* (*High Energy Physics*), this is the result of an international collaboration including countries from *ERA* (*European Research Area*), US, Japan and many others.

Its collider, the *LHC* (*Large Hadron Collider*) in Geneva, enabled experiments to obtain remarkable results on the verificating of the standard model, like Highs boson discovery in 2012. It provided to 10000 *LHC* users (7000 scientists and engineers among them) proton-proton collisions of 7 TeV of energy in centre-of-mass. Through the years this energy was increased reaching 8 TeV in centre-of-mass in 2013.

The future interest of *CERN* is to go beyond the standard model that is to verify

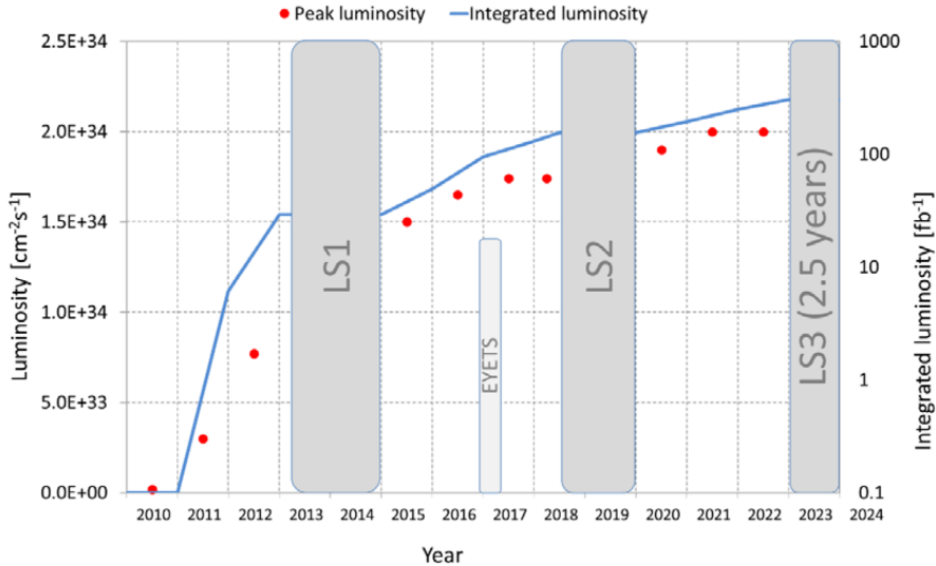


Figure 1.1: Planned *LHC* luminosity upgrades thorough the next years, taken from [1]

or disprove supersimmetric theories or search to acquire proof on the existence of dark matter. In order to fulfill these goals a major upgrade is needed: an increase of the luminosity, and thus of the collisions by a factor of five beyond its design value. For this reason the upgrade project of *LHC* is called *HL-LHC* (*High Luminosity-LHC*) [1].

With the usual upgrade programs, running the accelerator will become marginal due to statistical error: 10 years of data acquisition would be required only to have the statistical error. For this reason *HL-LHC* would pose the ambitious goal of achieving a 3000 fb^{-1} of integrated luminosity in a dozen of years, all at the nominal energy of 13-14 TeV in centre-of-mass, which corresponds to $\times 10$ of the design integrated luminosity value. But, in order to reach 3000 fb^{-1} with a 50% margin in effective data acquisition in full operation, the peak luminosity needs to be pushed to $7 \div 7.5 \times 10^{34} cm^{-2}s^{-1}$. The upgrade program of the next years is presented in fig.1.1.

Reaching this goal would mean to fully exploit the *LHC* facility, it requires $300 \div 350$ fb^{-1} per year, which corresponds to a peak luminosity of $5 \times 10^{34} cm^{-2}s^{-1}$. The project started in 2010 and it entered in the nominal energy regime in 2015, with a luminosity of $1 \times 10^{34} cm^{-2}s^{-1}$ (40 fb^{-1} per year). The timetable foresees the end of the R&D in 2025 and it will have to assure high efficiency in operation until 2035. The first upgrade of the luminosity is planned for 2022 and it aims to double the 2015 value. Experiments need to be upgraded in the mean time in order

to be ready for *HL-LHC*, the upgrade will range from the experiments structure to sensors, data acquisition, process and electronics.

The development of electronics for *LHC* experiments started thirty years ago and it has enabled the growth of the research field both in capacity and expertise. LHC has demanded through the years the development of many *ASICs* (*Application Specific Integrated Circuits*) with deep sub-micron *CMOS* technology (detailed in 2.2) as the baseline. *CMOS* offers good digital and analog performance and radiation hardness exceeding the requirements of the time (100 kGrey of *total ionizing dose (TID)* and $10^{14} MeV n_{eq} cm^{-2}$). This technology enabled the building of the readout of high density sensor systems with low power consumption and with short connections to the sensor.

The biggest challenge for the new electronics comes from the high event rate [2]: the total foreseen pileup (number of events per bunch crossing) in the detectors will grow from 27 to 200 (the initial design value was 19). Since the bunches cross every 25ns, the expected hit rate on the inner most sensor layer will go from the present $200MHzcm^{-2}$ to $3GHzcm^{-2}$, leading to an average rate of $75KHz$ for a pixel of $(50 \times 50)\mu m^2$ area. To prevent signals pile-up and provide a good particle tracking, timing informations are required. To assure good background rejection timing resolution needs to be under 1ns, this can be provided by triggers signals generated by electromagnetic calorimeters, which feature sub-200 ps resolution. The expected trigger rate will be of $1MHz$, requiring a data bandwidth of 1.25 $Gbits^{-1}cm^{-2}$. This performance could be met by processing the trigger matching close to the front-end in order to reduce latency and by increasing the buffers in order to accommodate it. This high precision timing standard is supported also by advancement in the sensors side: new 3D-Sensors features high radiation hardness and fast and reliable signals.

New possibilities both for integrating more features and for introducing timing measure capability on pixel area are opened by the adoption of new scaled technologies below 130nm and 65nm. This new technology comes with some drawbacks and challenges: the typical scaling Dennard scaling method which consists in scaling by a certain factor both voltages and transistor size has reached its limit (discussed in 2.3). This can be seen by observing the scaling in supply voltages from 1.2V to 0.7V despite a minimum feature shrinking from 130nm to 14nm: with usual scaling methods the supply voltage would have been scaled to 130 mV instead. In order keep on the trend of a 0.7x geometrical scaling every two years many new solutions have been employed resulting in more complications in *ASIC* design.

In any case, higher densities will allow designers to make more features in pixel

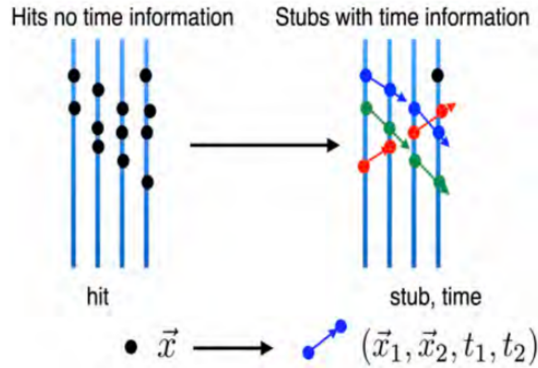


Figure 1.2: Conceptual representation of 4D tracking, taken from [3]

device front-end like digital calibration circuit for analog parts or *ToT* (*Time over Threshold*) measures for both timing correction and charge extraction. In terms of timing performance, the time measure resolution could be brought to sub-100ps by using *All Digital TDC (Time to Digital Converters)*, taking advantage of the good scalability of digital circuits. There are already examples of *TDCs* with 20ps resolution and 1mW of power consumption at $1M_{smaple}s^{-1}$ in 130 nm, new technologies could open the possibility of integrating one of them per pixel.

The critical factor in the adoption of all this new technologies is the *NRE costs (Non Recurrent Engineering)* associated with the masks (more on the manufacturing process in 2.2.2), right now 28 nm processes are becoming cost effective and will represent the target technologies for electronics for pixel sensor detectors.

1.2 The TIMESPOT Project

The TIMESPOT (*TIME and SPace real-time Operating Tracker*) project by INFN aims to overcome the *HL-LHC* challenges with a systemic approach, it proposes to research and develop a demonstration prototype of a complete 4D tracking system with $(55 \times 55)\mu m^2$ space resolution and at least 100 ps time resolution [3].

The 200 event pileup in future *HL* experiments is expected to increase the probability of ghost-tracks (from spurious hits) from 1.6% to 40%, thus damaging total tracking efficiency which is expected to reduce from 99% to 96%. For this reason high granularity and time resolution are essential to restore the tracking efficiency: the tracking algorithm will exploit *stubs*, namely timing track at detector level (illustrated in fig.1.2). This process is essential for background rejection, but in order to make it effective both sensor and front-end would need to be fast and precise, while

the tracking system needs to receive and process Tbs^{-1} of data. For these reasons the proposed solution is conceived as a system: every block needs to be optimized in conjunction with the others.

The system will be composed of:

– A high speed, high granularity and rad-hard 3D solid-state pixel sensor. The pixel pitch would be of $(55 \times 55)\mu m^2$ with resolution inferior to 100 ps. Two variants would be investigated: a silicon 3D sensors and diamond based one. Details on these sensors are presented in 1.2.1.

– A 28 nm *CMOS* front-end with timing measure capability. A dedicated analog chain will be designed for each sensor, considering the fluctuations that this step will add to signal, its design must be done in interplay with the ones of the sensors. Ideally every pixel will have its own 100 ps *TDC*. Every pixel will even integrate the digitization and eventual digital signal processing circuits.

– A real-time 4D tracking board: it will use a custom parallel track processor based on pattern matching with stored pre-calculated events in order to avoid combinatorial means. The board could implement this using FPGAs. An associative memory banks layer could be used as a first step producing a veto condition that will reduce the data bandwidth that needs to be processed. The process needs to be in real-time in compared to the 40 MHz bunch-crossing and it must be able to process 1 Tbs^{-1} of data.

To reflect this characteristic, the project is divided in 6 work packages (WP) across 10 INFN centers, it aims to develop in conjunction this 4D tracking system:

(WP1) 3D silicon sensor design and characterization

(WP2) 3D diamond sensor design and characterization

(WP3) front-end design with timing capability

(WP4&5) high speed read-out boards for real-time tracks reconstruction

(WP6) assembly of a working small telescope prototype called *Timespotter*, featuring at least 4 fully equipped tracking layers, high speed data tracking and real-time processing.

The research and development of the various elementary blocks will start as an optimization process of consolidated technologies, where the final goal is to produce

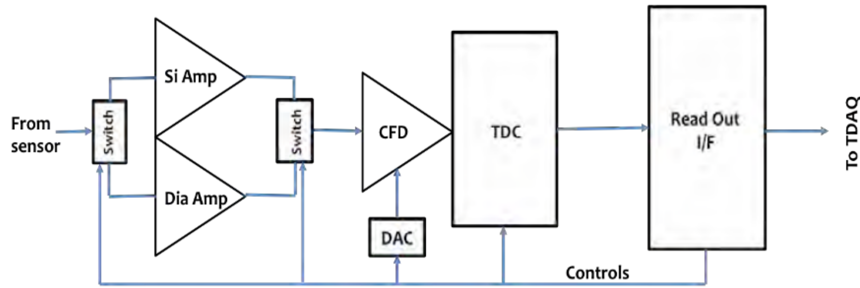


Figure 1.3: Block-diagram representation of the tentative front-end architecture, taken from [3]

a scale down prototype to demonstrate the feasibility of the system. The prototype will be scaled down in the number of channels but not in its features: it will be a complete tracking system. The first phase of the research will be focused on the first three WP and it will produce and study both the two type of sensors and essential elementary blocks of the front-end. This preliminary work will end in 2018 with the individual test, characterizations and debugging of these blocks. In the second phase in 2019 a fairly large pixel matrix, fully integrated, bonded with the sensor will be studied. A tentative architecture of the first part of the front-end is represented in fig.1.3. The first amplification stage could be exclusive for the two sensors due to their intrinsic differences (sensor capacitance, SNR, signal amplitude). The plan is to integrate both core amplifiers on pixel. Two Architectures were investigated: a fast but noisy amplifier-shaper architecture or a slower but with less noise *Charge Sensitive Amplifier (CSA)* architecture. The discrimination stage, essential to produce the digital signals triggering the timing measure, must be compact while not degrading the performance because of time-walk: a *Constant Fraction Discriminator (CFD)* is a possibility along with a *Leading Edge Discriminator* with *ToT* correction (details in 1.3). Finally an All Digital TDC will digitize the timing difference between the signal and the reference *LHC* bunch-cross clock with at least 100 ps LSB. The digital architecture is selected due to its optimal scaling to modern digital-oriented technologies. In case integrating a per pixel *TDC* will results unfeasible, the possibility of shared *TDCs* among 4 or 8 adjacent pixels will be explored. Assuming a natural scaling of a previous 130nm synthesis-able 200 ps architecture the area occupation could be of $(20 \times 30)\mu m^2$. In the end all this blocks will reside in the $(55 \times 55)\mu m^2$ pixel area.

The subject of this work is the very first analog part of this front-end, these structures are among the one that will be manufactured and individually tested

within the end of 2018 in a test chip. The selected amplifier architecture is the *CSA* one, as for the discriminator a compact leading edge one is initially studied in order to evaluate the discrimination performance, the impact of time-walk and the feasibility of *ToT* correction. This architecture will be discussed in chapter 3.

The implementation of a front-end with per-pixel sub-100 ps timing measure capability represents a novelty in *HEP* experiments: the best results in timing performance are represented by: the *TDCpix* for the *NA62 Giga Tracker* which features a $300\mu m$ pitch with 150 ps resolution obtained by multiplexing one 75 ps *TDC* for 1800 channels (not tailored for *HL*) in a 130nm technology node; and by *Timepix3*, developed at *CERN*, with a pitch of $55\mu m$, 1ns resolution with one *TDC* per 8 channels, in 130 nm.

1.2.1 TIMESPOT Radiation Sensors

Radiation sensors convert part of the energy deposited by an incident particle in an electrical signal. In the case of considered family of detectors, the signal is a current pulse whose integrated charge is the one created inside the bulk of a semiconductor thorough ionization. The development of the current signal is due to the induction on a metallic electrode caused by the motion of the mobile charge carriers created during ionization, namely electrons and holes. This is the drifting motion under the effect of the electrical field imposed by a bias voltage applied to a biasing electrode. The instantaneous current I_k induced on an electrode by a charge q can be expressed using the *Ramo Theorem* [4]:

$$I_k = -q\vec{v_d} \cdot \vec{E_w}(\vec{x}) \tag{1.1}$$

Where $\vec{E_w}(\vec{x})$ is called *weighting field* and it corresponds to the field generated by setting the potential of the specific electrode to 1 and the one of all others to 0. This field defines only how the charge couples with the electrode and it depends only on the geometry of the system. Charge velocity, and thus its trajectory, depends on the actual electrical field $\vec{E}$ thorough the *drift velocity* v_d:

$$v_d = \mu E \tag{1.2}$$

Where μ is the carrier mobility. This is true in case of velocity saturation: electrical field needs to be sufficiently intense to satisfy this condition (typical values exceed $10^4 Vcm^{-1}$). It should be noted that $\frac{\vec{E}(\vec{x})}{|E(\vec{x})|} = \vec{E_w}(\vec{x})$ only in two-electrodes systems. Both type of carries contribute to the total current in the same direction since both

their charges are opposite in sign but at the same time they will drift in opposite directions. The duration of the signal is the one required to collect all the carriers inside the metallic electrodes: once inside the metal volume, the total charge induced by them equals zero.

The first consequence of this fact is that the shape of the current pulse depends on the shape of the charge development inside the volume: electrons and holes contribution will last as long as they reach the respective electrode which is given by their distance from them and their drift velocity (note that electrons and holes have different mobility). This constitutes an uncertainty on the timing of the signal.

There are other sources of fluctuation of the signal: the number of pairs crated during the particle interaction, the possibility of carriers trapping due to impurities on the material and the delta-ray effect (production of a secondary electron far from the primary interaction which can itself ionize the material).

One last cause of uncertainty is the eventual presence of low-field zones which makes the velocity saturation condition no more true.

The adoption of 3D sensors for timing detectors is driven by their capability to mitigate most of these effects and thus to produce more reproducible signals: standard 2D sensors present the electrodes arranged on the surface of the material, while in the 3D ones they penetrate inside the bulk volume. This sensor geometry, compared to the flat one, makes it possible to put the electrodes close to each other (inter-electrodes distances $< 70\mu m$) without reducing the sensitive area, preserving the sensor efficiency and improving the charge development. With depth $> 200\mu m$ more electron-hole pairs can be created, thus increasing the signal strength. Furthermore the shorter electrodes distance allows to reach the desired electrical field strength with a smaller bias voltage.

In terms of signals reproducibility, electrodes proximity brings many advantages:

- for perpendicularly incident particles the absolute signals length variation for the opposite extreme cases (adjacent to the electrode of one polarity versus the other) is reduced.
- the inter-electrodes distance (L_e) could be made small compared to the carriers mean free path due to trapping, reducing the overall trapping probability.
- L_e can be even made shorter than the delta-ray creation distance, eliminating the contribution from this effect

The higher presence of low-field zones due to border effect caused by the tips of the electrodes (usually column like shaped) is the main disadvantage in terms of time

resolution of 3D sensors. The current pulses from these cases presents a long decay tail. However this contribution can be minimized acting on the electrodes geometry: reducing the relative low-field volume making the charge formation in that zone less frequent or making the field variations less prominent in order to make the timing difference lower.

Another advantage of 3D sensors which makes them preferred solution for *HL* experiments is their intrinsic radiation hardness: the principal radiation damage is the creation of defects inside the crystal which act as traps, both the lower required bias voltage and the short inter-electrodes distances mitigate the contributions.

However 3D sensors present generally more capacitance since electrodes and the sensible area forms a capacitor with small distance between plates. As a consequence, the sensor-amplifier system becomes slower and with a smaller SNR.

As a consequence of these aspects, in order to obtain the desired resolution of 100 ps, front-end and sensor must be optimized together: solutions for the sensor signals fluctuation can introduce a problem for the analog signal processing on the very first-end, trade-offs between the two part must be considered. As detailed in chapter 4, in order to tackle this need, simulations during design phase of the front-end were performed on signals derived from simulations of the sensors, enabling, when necessary, work in progress adjustment to both blocks.

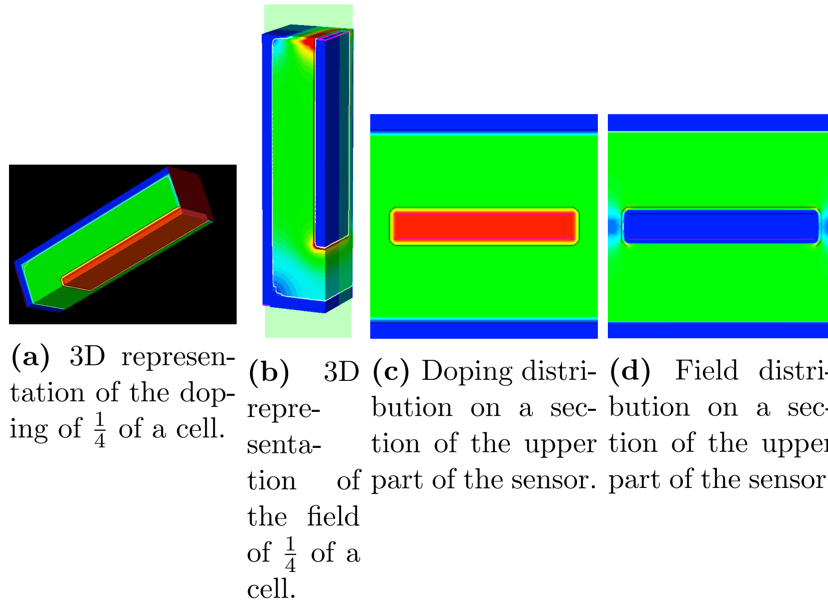
In the following sections characteristics of both 3D sensors realization investigated for *TIMESPOT* project are detailed.

Silicon 3D Sensor

The silicon based sensor will be designed by INFN Trento [5]. Its dimension are $(55 \times 55 \times 100)\mu m^3$ with a trench shaped $(6 \times 45 \times 75)\mu m^3$ central signal electrode and $6\mu m$ wide bias electrodes, parallel to the signals ones, which separates the cells along one axes (presented in fig.1.4a and fig.1.4c). The substrate is p^{--} doped (weakly doped with acceptors), while the signal electrodes are n^{++} (highly doped with donors) and the bias ones are p^{++} (highly doped with acceptors). At the bottom of cells $3\mu m$ p^{++} layer. The bulk silicon is p^{--} due to its radiation-resistance properties.

Many geometries were tested, the selected one features the most uniform field obtained (represented in fig.1.4b and fig.1.4d): there is a low filed area along two adjacent n^{++} electrodes which correspond to the 1.5% of the area of a section of the electrode (fig.1.4d); another low-field is at the bottom of the cell under the tip of the electrode(fig.1.4b), this accounts for the 10% of the total 3D volume.

The manufacturing process used consists in etching the silicon bulk with *Deep*



(a) 3D representation of the doping of $\frac{1}{4}$ of a cell.

(b) 3D representation of the field of $\frac{1}{4}$ of a cell.

(c) Doping distribution on a section of the upper part of the sensor.

(d) Field distribution on a section of the upper part of the sensor.

Figure 1.4: 3D silicon sensor dimensions. The doping is represented from red(n^{++}) tu blue(p^{++}). The field is modulus is represented with blue to green for low field to high field values, biased with 100V.

Reactive Ion Etching (DRIE) and then doping the holes with desired atoms. These holes are then filled with polysilicon in order to create the ohmic contacts, this process is the one limiting the smallest electrode that can be produced: filling reliably these holes become difficult the smaller their section. At the same time making electrodes excessively large will reduce the active area, reducing the efficiency. To use this process a support wafer is needed: a SiO_2 layer is used and than removed with a further step. On this back surface the bottom p^{++} doping is realized: since the bias electrodes passes through the whole bulk layer reaching the previously removed SiO_2 layer, the bias voltage can be directly applied to the newly formed p^{++} layer. Attention has been kept to the distance between the p^{++} layer and the n^{++} electrodes since a too thinner ones cause low sensor break-down voltages.

The sensor thickness, and thus the electrodes length, is limited by two factors: the sensor capacitance C_d increases with the electrodes depth and the poly filling becomes less reliable. In any case due to the shape of this sensor resembling the one of a parallel plate capacitor its C_d is fairly large: $\sim 100 fF$ (including the front-end connections).

The front-end connection will be made my bump bonding directly on the top surface of the sensor.

Simulations of this sensor were performed by INFN Cagliari, providing informa-

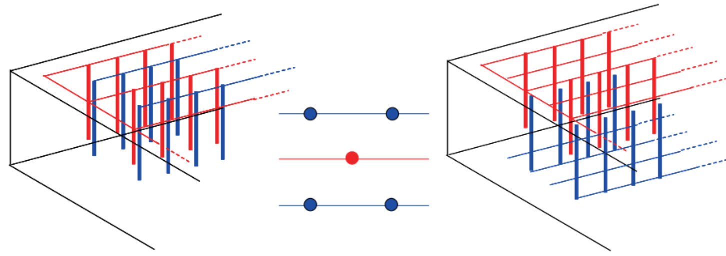


Figure 1.5: Schematic representation of the electrodes positions in the diamond sensor for both one and two-sided variations, taken from [6]

tion on field distribution and on the expected signal. The simulation set up was: bias voltage set to 100V and uniform 1.2 fC linear charge development perpendicular to the top surface. The expected signals are expected to have a current peak of $\sim 10\mu A$ and duration $< 300ps$. There are plans for future analysis on different angles of incidence and non uniform charge distributions calculated with a radiation-mater interaction software (GEANT4 from *CERN*).

Diamond 3D Sensor

The 3D diamond sensor is designed by INFN Firenze [6][7] and modelled and characterized by INFN Perugia.

It consists in graphite electrodes inside a CVD (*Chemical Vapor Deposition*) made polycrystalline diamond. The bulk is made with hetertoepitaxial growth on top of iridium, it is a low cost solution enabling the formation of samples with a $(2 \times 2)cm^2$ area and 500 μm thickness.

The columns shaped electrodes are made by direct graphitization of the diamond using pulsed laser along the beam focus, a technique originally used for optical application. To trigger the transformation process, the laser energy density per pulse must exceed $5Jcm^{-2}$. Columns shaped electrodes with a diameter of 9 to 10 μm can be produced. Two different lasers were employed: a 8ns Nd:YAG, the other 30fs Ti:sapphire. Different results are met depending on the laser used: ns one has low collection efficiency but low resistivity, while the fs one has higher resistivity ($60k\Omega$ versus $4k\Omega$). Raman spectroscopy suggests the presence of sp3 bonds on the electrodes surface as an explanation of ns low efficiency. Multiple steps gratifications can be employed to enhance the electrodes properties, but it is time consuming.

The diamond is naturally a good material in terms of radiation hardness, and it seems to improve after irradiation. It even presents high uniformity, and is polarized with 100V bias. The resulting signal is weaker than the one of silicon detectors, bu

the sensor presents better SNR and lower capacitance.

The fabrication technique allows to freely trace 3D structures, the one proposed is formed of hexagonal combs in which the center and vertex electrodes are oppositely polarized (fig.1.5). This creates an interdigitated structure in which the electrodes can be connected to the surface by creating paths to the surface. Connection can be made on either of two top and bottom surfaces opening the possibilities for one or two-sided sensors, however the actual bonding mechanism with silicon is yet to be decided: the two solutions are direct diamond-silicon connection using the laser or metallic bump bonding.

1.3 Electronics for Timing

A timing measure is a value proportional to a timing interval, in most of cases this is the time elapsed between a reference event and the measured one. This section will present the main methods used to extract a timing measure from an electrical signal in order to convert them in a digital number. First, in 1.3.1, some useful timing metrics used in electronics are presented. In 1.3.2 some electronics architectures are illustrated, their feasibility in the context of *CMOS* technology for pixel sensors is also discussed.

1.3.1 Timing Metrics

Single Sample

The time information can be taken from sampling the signal in a predefined condition of its value. A typical condition is a threshold crossing in a certain direction. The absolute time information is the elapsed time from a well defined event, which can be measured in the same way. Good reference signal must be periodic, stable and will allow a precise measurement of the condition event. An example which is often found in electronics is using a steep rectangular-shaped digital signal with a certain periodicity, where the crossing of a certain value of its steep rising or falling edge is used as the time reference.

The simplest example of method which provides the extraction of timing event from an analog signal is the *Leading Edge Discriminator*. With this method a digital pulse is produced with its value (ideally) instantaneously transitioning on the first threshold crossing of the signal. Threshold value and the direction of its crossing must be decided based on the signal characteristics like polarity, base line value and noise.

As for the noise, the statistics of the signal must be even taken into account: in order to avoid false crossing-event the threshold must be put at level far from the noise one. In case of normal distributed noise, the threshold must be kept at a level which makes the rate of spurious crossing-event below the desired value. In any case the noise is superimposed to the signal, thus for real signals with finite non-instantaneous rising times, the signal leading edge will be disturbed creating the so called *jitter* (σ_t). Approximating linearly the rising profile, with a peak value V_p and rising time t_r:

$$\frac{dV}{dt} \sim \frac{V_p}{t_r} \tag{1.3}$$

And considering a noise voltage amplitude σ_V, the jitter can be computed as:

$$\sigma_t = \frac{\sigma_V}{\frac{dV}{dt}} = \frac{t_r \sigma_V}{V} = \frac{t_r}{SNR} \tag{1.4}$$

When the desired event to be measured is the starting of the development of the signal on the base line, the threshold cannot be positioned close to the baseline due to noise triggering. Selecting the threshold to an higher value may not be an applicable solution because of signal nature: if the slope is not constant for every signal the delay time between this event and variation from the baseline will not be constant to. In case of the signal presenting constant rise time (typical of bandwidth constrained amplifiers often used in particle detectors as the *charge sensitive amplifiers*) but variable peak values, the threshold is crossed at a different time in respect to the amplitude of the signal. This undesired effect is called *time-walk* and is depicted in fig.1.6. Calling m the ratio between the amplitude of two signals:

$$t_c = \frac{V_{thr}}{d_t V} = \frac{V_{thr} \cdot t_r}{V_p} \Rightarrow V_{p2} = m V_{p1} \Rightarrow t_{c2} = \frac{t_{c1}}{m} \tag{1.5}$$

As it can be noted, the time-walk has an hyperbolic relation with the amplitude increase.

With these signals of this type, the jitter is related to the amplifier bandwidth (BW), in fact knowing that $BW \propto \frac{1}{t_r}$ and $SNR \propto \frac{1}{\sqrt{BW}}$:

$$\sigma_t \propto \frac{1}{\sqrt{BW}} \tag{1.6}$$

The time-walk constitutes a systematic error which, in case of it being relevant in magnitude, it must be accounted and corrected. If the signals amplitude is known

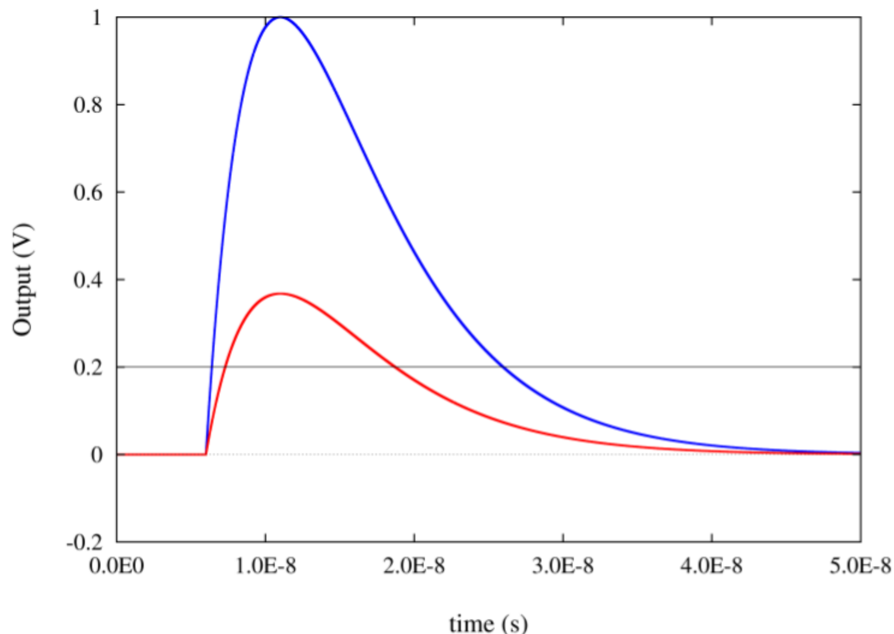


Figure 1.6: Example of time-walk on a signal with constant rising and falling times.

from other measures it will identify the time-walk contribution on the time measure. A method which, under certain assumptions on the signals shape, can uniquely identify the signal amplitude exploiting another time measure is the *Time over Threshold (ToT)* corrections. The ToT is the time in which the signals stays above (or below depending on its polarity) the threshold: two crossing events measures of opposite direction are needed. This value can be used to correct the time-walk only in the case the relation between V_p and ToT is known and unique. Assuming (as in the case of the proposed amplifier detailed in 3.2) that both the rising and falling time are constant, the relation between amplitude and ToT can be found. If threshold level is in a constant slope region in both crossing times the relation is expected to be linear.

Multiple Samples

The desired timing information of a signal can be extracted by sampling the slope of the signal at different times, if the shape of signal is known. Fig.1.7 depict this method. The advantage of this approach is that, for an infinite number of samples, it tends to cancel the random contribution of noise. Since the noise contribution to every sample is independent, sampling the signal N times will reduce the previously

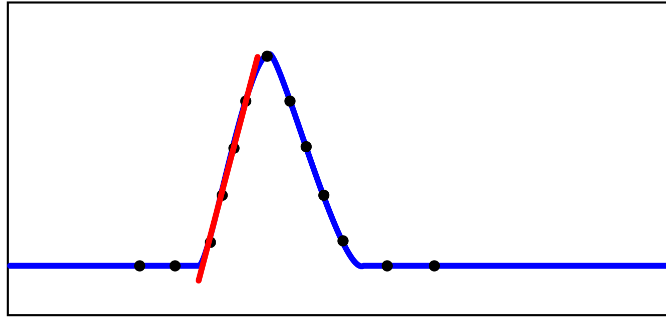


Figure 1.7: Example of multiple samples take on a signal

obtained relation by:

$$\sigma_t = \frac{t_r}{SNR}\frac{1}{\sqrt{N}} \tag{1.7}$$

This can be expressed by the sampling frequency $f_s = \frac{1}{t_s}$ knowing that the number of samples into the rising edge of the signal equals $N = \frac{t_r}{t_s}$:

$$\sigma_t = \frac{1}{SNR}\sqrt{\frac{0.35}{BW \cdot f_s}} = \frac{1}{SNR}\sqrt{\frac{1}{3 \cdot f_{-3dB} \cdot f_s}} \tag{1.8}$$

Where BW is the bandwidth of the amplifier, therefore oversampling is needed to reduce significantly the jitter.

Although, in order to obtain a high sampling frequency and measure, the signal value both a high frequency ADC is needed. This circuits takes large chip areas and will be really power hungry operating at the velocity requested for the application: the signal can't be made too long in order to avoid pile-up due to high event-rate. It results, from these considerations, that oversampling the signal is not a solution for timing pixel detectors for high luminosity HEP experiments.

1.3.2 Overview on Timing Architectures

In this section various architectures, which can be employed in this specific case, are briefly illustrated. This is a block-diagram level description and not a practical *CMOS* realizations. First the two main candidates as discriminators are discussed, then an overview on various *Time to Digital Converter (TDC)* is presented.

Leading Edge Discriminator

The intended operation of this circuit is discussed in the previous section. Fig.1.8 illustrates the block scheme of the simplest architecture and its ideal response. The depicted ideal behavior is compromised by the limited gain and bandwidth of the amplification stage: in fact only with infinite gain and BW a zero rising time for the output signal can be obtained. With a finite gain, the output signal slope will retain some information of the input one, making the delay time dependent on the signal slope in the proximity of the threshold. The BW limitation will, in case of infinite gain, make the delay time equal to the slew-rate of the amplification stage.

A leading edge discriminator is generally composed by a differential high gain amplification stage that sens the threshold crossing and amplifies resulting in a steep signal. It is then followed by a digital buffer which produces the desired digital signal.

The amplification stage can be composed of many lower gain stages for two principal reasons:

- Making a single high gain differential stage in *CMOS* is impractical, so a two stage configuration with a single ended high gain stage can be the preferred solution.
- Generally, in *CMOS* technologies, making a high stage gain will reduce its bandwidth creating a gain-bandwidth trade off. This situation can be avoided by making many low gain stages, sized so that the BW will be little compromised, but obtaining the same total gain. Therefore this method is used to obtain fast discriminators.

It must be said that fast discriminators, due to the big number of requested stages, tend to be large an power consuming. So they are not an option for the desired front-end.

In any case, this simple type of discriminator is used as a basic building block for more complex ones.

Constant Fraction Discriminator

As the name suggests, *Constant Fraction Discriminator (CFD)* is a discriminator which produces a signal when a threshold relative to the signal peak is crossed. What the user can do is to change value of this fraction of the total signal. The direct application of this discriminator is to prevent time-walk.

In order to make the process work, the assumption of a signal approximated to a triangle-shaped one, with constant rising and falling times in the entire signals fam-

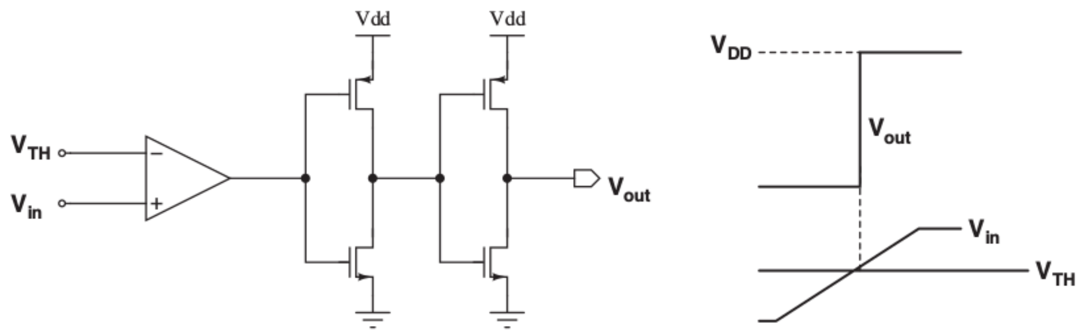


Figure 1.8: Concept of leading edge discriminator and ideal signal response, taken from [8]

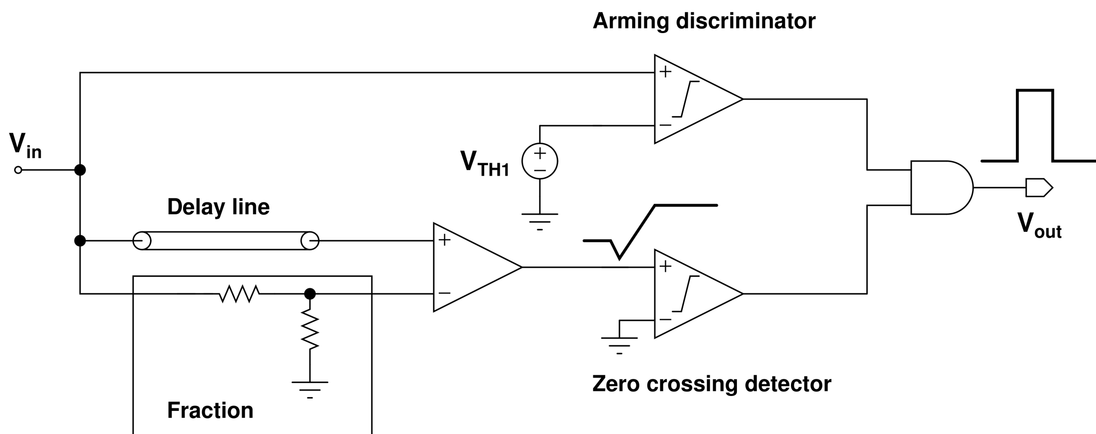


Figure 1.9: Concept of CFD, taken from [8]

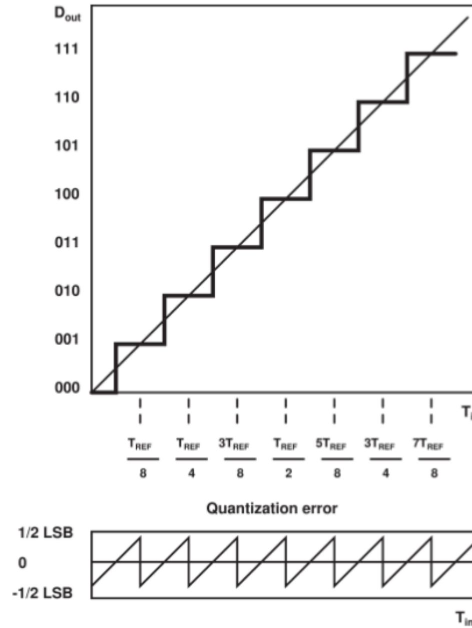


Figure 1.10: Ideal signal response and quantization error of a TDC, taken from [8]

ily, must be hold. Practically, the constant fraction is realized by firstly producing a bipolar signal from the input one: the signal is delayed and then subtracted with an attenuated version of itself:

$$out(t) = in(t-d) - f \cdot in(t) \tag{1.9}$$

Where d is the amount of which the signal is delayed and f is the attenuation factor. The zero crossing point of this signal happens when a certain fraction of the signal is reached, for every signal, so a zero crossing discriminator is used to trigger the condition. An arming discriminator is then used to vetoing the noise-trigger from this last discriminator. Fig.1.9 illustrates this architecture.

The fraction can be implemented with a simple resistive partitioner. As for the delay, it can be realized with passive filtered line [9]. Although a passive implementation is difficult to implement with deep-submicron *CMOS* technologies due to their limited voltage headroom (discussed in 2.3.5) the big area occupation of capacitors and resistors.

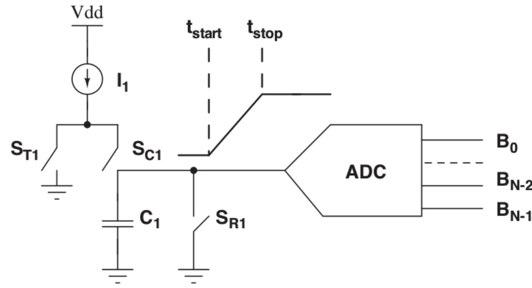


Figure 1.11: Concept of an Analog TDC using the voltage across a capacitor charged by constant current source as TAC, taken from [8]

Time To Digital Converter (*TDC*)

TDCs are a family of electronics device which directly digitize a timing interval. The input signal can be structured as a time displacement between two well defined steep signals, or a single rectangular signal which holds a certain state (ad example a voltage value) for the time period that will be measured.

Since a *TDC* is a digital converter it will produce a set of discrete values which the minimum one (represented by its LSB) defines the best resolution achievable by the system. The absolute timing value that corresponds to the LSB depends on the mechanism used to discretize the timing interval, the various *TDC* architectures differentiate one to the others on the base of it. As for the number of bits used, they are not upper limited, but they are usually constrained by circuit complexity, area usage, time required to complete the measurement and power consumption. For this reason the resolution can be expressed (for linear converters) by the largest timing interval measurable (range T_r) divided by 2^{N_b}, where N_b is the number of bits. In fig.1.10 the ideal signal response of a *TDC* with its quantization error is presented.

However non-linearities in the conversion procedure can reduce the resolution within the range. For this reason the average resolution of the *TDC* can be expressed by using its *effective number of bits ($ENOB \leq N_b$).* This number is reduced as well by the sensibility to the noise of the conversion. The non linearity of a system is usually represented by the *Differential Non-Linearity (DNL)*:

$$DNL = \max_{1<i<2^{N_b}} \left[\frac{t(i) - t(i-1)}{t_{LSB}} \right] \tag{1.10}$$

Where t_{LSB} is the ideal minimum time step. The DNL is usually presented in units of LSB. So, in order to create a high resolution *TDC*, a fast, linear and low noise sensitive technique needs to be exploited.

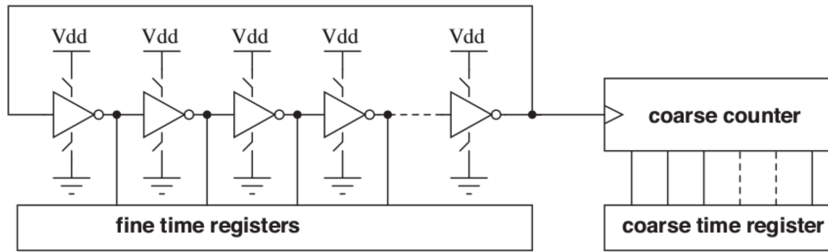


Figure 1.12: Example of gated ring oscillator, taken from [8]

There is a grate variety of *TDC* architecture, but mainly they are divided in two groups: *Analog TDC* and All Digital TDC. Devices of the first category digitize a time difference which was previously converted in an analog (voltage or current) signal with a *TAC (Time to Analog Converter)*, and the digitized with an ADC. Fig.1.11 shows an example of an Analog TDC: when the start signal arrives, S_{C1} is opened and S_{T1} is closed, in this way I_1 starts charging C_1; when the stop signal arrives the ADC starts converting the voltage; when the conversion is completed C_1 is discharged using S_{R1}. Resolution and conversion time depend on the ADC properties constituting the two major drawbacks of this family of *TDCs*.

The second *TDCs* category is the so called *Fully Digital TDC (FD-TDC)*, this type of *TDCs* directly digitize the time interval. A first idea could consists in simply counting a high frequency clock, but this approach has its shortcoming: even though $\sim 10GHz$ clocks can be reliably generated, distributing reliably on a multichannel chip could be cumbersome and will dissipate a lot of power.

For these reasons a simple approach is to generate a clock in loco using a *Ring Oscillator* (an oscillating circuit formed by an odd number of inverters), and using its frequency as a coarse counter, while the change of state of its inverters will trigger the sampling of the status of the signal to be measured. The ring oscillator can be "gated" (making it oscillating when desired) in order to save power, the gate oscillator concept is shown in fig.1.12. These types of TDC are generally called RO-TDC, and their resolution is limited by the minimum delay inverter that can be produced in the selected technology.

Another approach is the one of the *DL-TDC*s (showed in fig.1.13), based on delay lines. The simplest topology uses two rising (or two falling) edges of a signal which was previously at 0. The start signal is then delayed using voltage buffers, the stop signal is then used as the dynamic signal of DFF based registers which samples the delay line status. When the stop signal rising edge comes before the

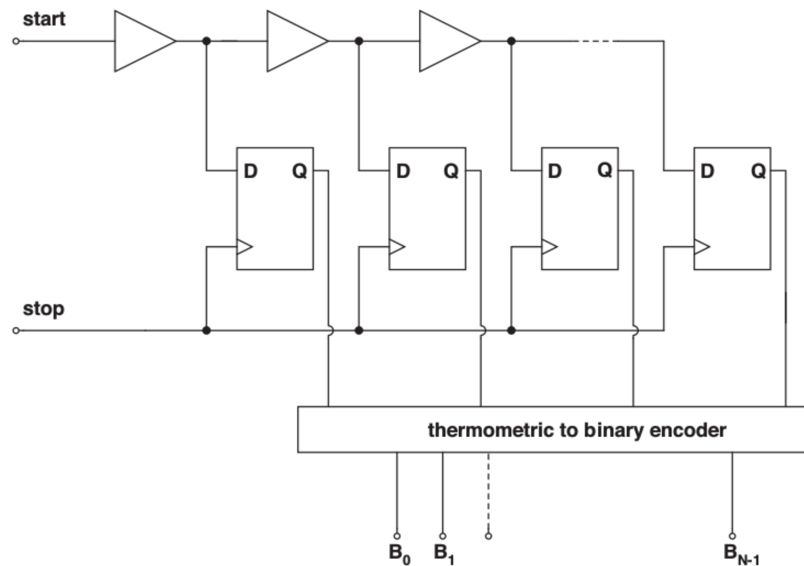


Figure 1.13: Delay line based TDC, taken from [8]

start one, a 0 state is sampled and the measure ends. The start and stop signals must be re-setted to 0 afterward.

In this way the range depends on the length of the line resulting in a waste of chip area, in order to prevent this the delay line can be looped, counting the number of the elapsed loop will provide a coarse counter. Another modification of this topology enables to obtain an inferiori resolution to the minimum gate delay: with the *Vernier* topology (fig.1.14 both signals can be delayed by a different amount resulting in a resolution equal to the difference of the two delay times.

A last topology is the one based on *Pulse Shrinking* (fig.1.15). As the name suggests, the signal is progressively time shrieked and used as dynamic signal of a DFF to the point its duration becomes so small that no more can be detected. Even this topology can be looped as the previous one.

Everyone of these topologies can be combined with a *Time Amplifier*, which stretches the timing difference between two signals, this enables to obtain high resolution using low resolution topologies at the expanse of linearity and conversion time. Two or more architectures with different resolution can be putted in stages: the less precise would take a coarse measurement, while the time difference between its measure and the signal can be measured by the more precise stage.

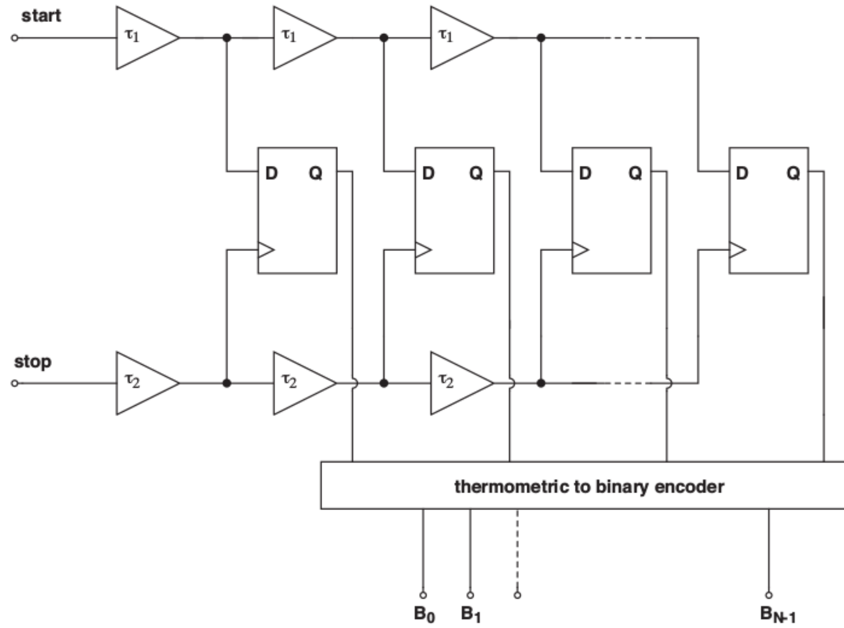


Figure 1.14: Vernier TDC, taken from [8]

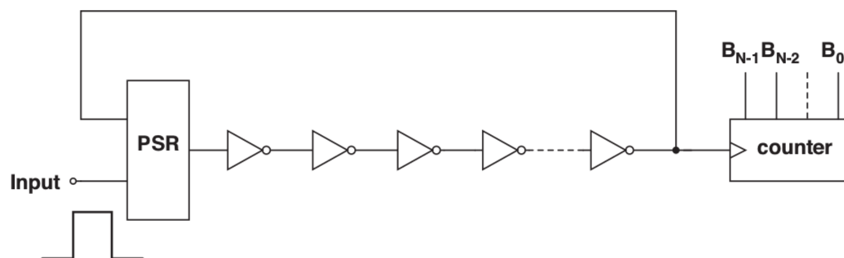


Figure 1.15: Pulse Shrinking TDC, taken from [8]

Chapter 2

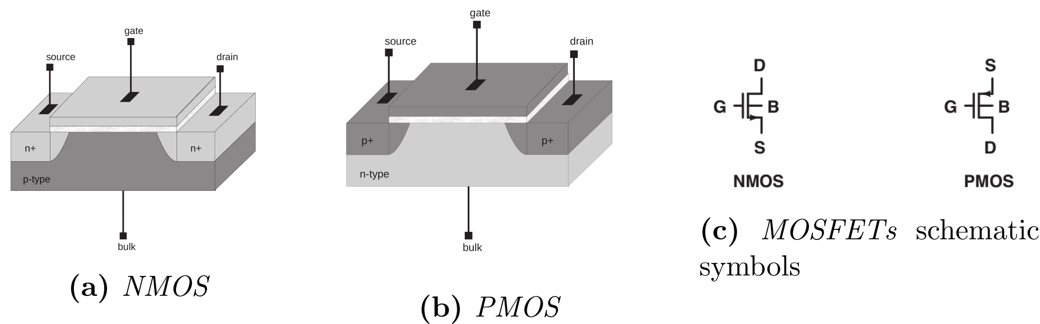
Deep Sub-Micron *CMOS* Electronics

This chapter aims to introduce key the aspects of analog *CMOS* electronics through its: operation, manufacturing process and design. These concepts will be later used to describe and discuss this work.
Section 2.1 is a brief recap on basic electrical characteristics of *MOSFET*'s in the designed front-ed: this is far from a complete treatise on the subject but is meant to introduce concepts to be used later.
In section 2.2 a number of technical aspects of *CMOS* including its manufacturing and design are illustrated. Section 2.3 focuses on the scaling to the 28nm node; the consequences on circuits design is also discussed.
The end of the chapter, section 2.4, is dedicated to an overview of various *CMOS* structures used in the actual front-end, including their mathematical small signal models.

2.1 *MOSFET* Properties

Metal Oxide Semiconductor Field Effect Transistors (MOSFET) are one of the most used and produced electronics devices in modern electronics. Their physical working principles are well known and will not be explained in this work [8], instead this section will include a brief presentation and discussion that aims to establish common terminology.

A *MOS* transistor is a device in which the current flowing in the *channel* between it's two terminals named *drain (D)* and *source (S)* is controlled by the voltage applied to its *gate* terminal (G) referred to the S terminal (V_{GS}). There are two types of



(a) *NMOS*

(b) *PMOS*

(c) *MOSFETs* schematic symbols

Figure 2.1: Taken from [8]

MOS transistors depending on the relation between the controlling voltage and the channel conductance, in fig.2.1c the schematic symbols for both are illustrated. In a *N-type MOSFET (NMOS)* the channel *transconductance* increases increasing V_{GS}, in a *P-type MOSFET (PMOS)* it increases with a negative gate-source voltage.

In fig:2.1a a representation of an *NMOS* device is presented. Here we can see that in *NMOS* the transistor *bulk* is made of silicon doped with *acceptor* atoms (called *p doping*, with a concentration of $N_A \sim 10^{16} cm^{-3}$). On this substrate two regions highly doped with *donors* (n^+ doping, with $N_D \sim 10^{19} cm^{-3}$) which forms the *drain* and *source* of the *NMOS* are implanted . The *channel* is the zone between the two diffusions and it is insulated from the gate with an insulator usually made of *silicon dioxide* (SiO_2). The *gate* itself is made of *polysilicon*: a ploycristalline state of silicon, selected due to its chemical affinity to SiO_2. This part is also doped n^+. A fourth contact is placed on the *bulk* (B) in order to control its voltage.

When $V_{GS} < 0$ the holes present inside the p-type bulk starts to migrate underneath the insulator in order to mirror the charge in the gate.
In this condition no current can flow since they form two reverse-biased junctions with the *bulk* (*accumulation region*). When V_{GS} raises the holes are repelled from the *channel* leaving only the fixed acceptor ions which cannot contribute to the current. When a certain threshold voltage (V_{th}) is reached, an electrons layer (*inversion layer*) forms inside the channel, this charges can now drift under the effect the field imposed by V_{DS} creating the *drain source* current (I_{DS}). The same but reversed reasoning applies for the *PMOS*: the differences are that the channel carriers are holes instead of electrons and that V_{th} is negative, a representation of this transistor can be found in fig.2.1b.

MOSFET electrical characteristics are modelled on the basis of its geometrical structure and the physical properties of the materials. In particular the two main physical parameters that affect the *MOSFET* operation are the *carriers mobility*

($\mu_n \sim 1500cm^2/V \cdot s$ and $\mu_p \sim 500cm^2/V \cdot s$, for electrons and holes in intrinsic silicon) and the *insulator dielectric constant* ($\varepsilon_{ox} = 3.45 \cdot 10^{-2} fF/nm$ in case of SiO_2). The most important geometrical parameters are related to the *channel* extension (width W and length L) and the *insulator thickness* (t_{ox}, typically few nanometres). It should be noted that with L will be indicated *effective length* of the *channel*, this is the length of the insulator layer minus the portions of the diffusions that extend underneath it.

Some of this parameters are technology dependent: μ_n and μ_p depend on the concentration of dopants on silicon, ε_{ox} depends on the selected insulating material and t_{ox} is chosen by the manufacturer in order to avoid leakage current through the insulator.
W and L, instead, can be tuned by the designer, although their values are bounded to be inside the maximum and minimum values ranges provided by the manufacturer. Of this rules the *minimum channel length* L_{min} is the one defining the technology node; the shorter this parameter is the more advanced is the technology as it allows the integration of more transistors on the same chip area.

The last aspect of the technology that must be taken into account is the fact that the *MOSFET* behavior depends on the voltages applied across its terminal; this identifies a number of working regions with their characteristics.

Through out the section electrical characteristics, small signal parameters and sources of noise of *MOS* transistors will be presented.

2.1.1 Electrical Characteristics and Working Regions

From an electrical point of view, the *MOSFET* operation can be divided in three main regions based on its I_{DS} current.
When $V_{GS} - V_{th} \leq 0$ the device is in *depletion*: there are no free charge carriers on the channel thus no current can flow between D and S.
When $V_{GS} - V_{th} > 0$ and $V_{DS} \leq V_{GS} - V_{th}$ the device is in *linear region*: the transistor works as a voltage controlled linear resistor.
Increasing V_{DS} moves the transistor in *saturation*: here, since the maximum current is limited, the transistor operates as a voltage controlled current source. Moreover there are three sub regions based on the level of inversion of the channel: *strong, moderate* and *weak inversion*.
A bands diagram of the gate-insulator-channel system sowing the bands bending for different bias conditions is presented in fig.2.2.
The relation for *MOSFET* capacitance and V_{th} are presented in first place in order

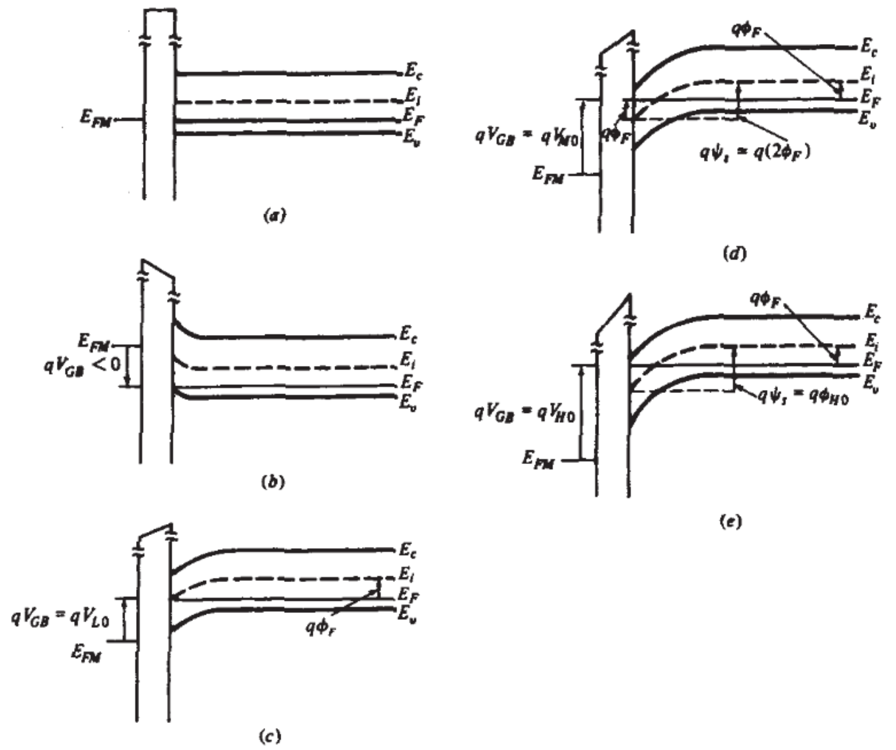


Figure 2.2: Band diagrams showing bands banding of *NMOS* in various bias conditions: (a)flat-band (b)accumulation (c)weak inversion (d)moderate inversion (e)strong inversion. The three materials are respectively the ones of gate-insulator-substrate (with metallic gate). E_c, E_v, E_i and E_F are the energy levels of conduction band, valence band, intrinsic silicon Fermi level and the junction Fermi level. It can be noted how, from accumulation to strong inversion, the Fermi level is moved from the valence band to the conduction one. In case of a polysilicon gate, the band bending will occur even inside it. Taken from [10]

to evaluate I_{DS} relations for all the operating regions.

Gate Capacitance

An important quantity used in further calculations is the capacitance per unit area (C_{ox}) of the insulator layer. In a first approximation it can be calculated as the capacitance of parallel plate capacitor:

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} \tag{2.1}$$

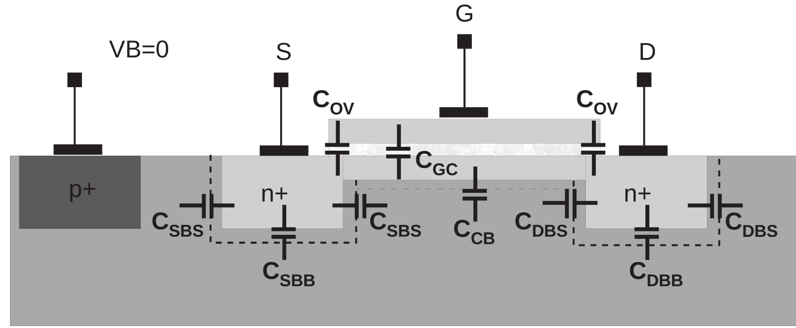


Figure 2.3: graphical representation of *MOSFET* capacitances, taken from [8]

The first useful characteristics derived from this quantity is the *gate-channel* capacitance, calculated using the same approximation:

$$C_{GG} = C_{ox}WL \tag{2.2}$$

Using this value, a first evaluation of the capacitance between the *gate* and the *drain* or *source* terminals can be calculated as:

$$C_{GS} = C_{GD} = \frac{C_{GG}}{2} \tag{2.3}$$

This capacitances are the one slowing down the *MOSFET* therefore limiting the bandwidth of its transfer function. For a more precise evaluation on the *MOSFET* capacitances, more effects need to be accounted: for every junction inside the transistor creates a parallel face plain capacitor, the thickness of this parasitic element depends on the extension of its depletion region: it varies dynamically with the bias. Fig.2.3 represents the other parasitics capacitances.

Threshold Voltage

As previously discussed this is the voltage required to form the *channel* inside the transistor. The value of this parameter depends on the device's physical properties and on the potential applied to the *bulk* terminal. For a first evaluation its value with *bulk* and *source* at the same potential is presented:

$$V_{th0} = V_{fb} + 2\phi_F + \frac{1}{C_{ox}}\sqrt{2\varepsilon_{Si}qN_A2\phi_F} \tag{2.4}$$

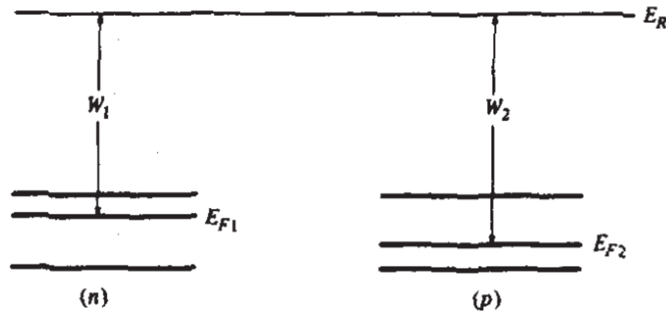


Figure 2.4: Band diagrams for the work functions of both *NMOS* and *PMOS*, taken from [10]

Where the term $2\phi_F$ is the voltage required to neutralized the unloaded *bulk*: its value depends on the *Fermi potential* of the substrate [1] , therefore V_{th} depends on the temperature. The last term accounts for the voltage generated by the exposed charge of the depletion region of temporarily created junction between the electrons *channel* and the *p bulk*.

V_{fb} is called *flat-band voltage*. It is the voltage required to avoid the bending of the bands inside the materials and therefore needed to obtain a uniform charge distribution. This condition of flat-band is depicted in 2.2(a), V_{fb} follows the relation:

$$V_{fb} = \phi_G - \phi_B - \frac{Q_{ox}}{C_{ox}} \tag{2.6}$$

Where ϕ_B and ϕ_G are the *contact potentials*[2] of the *bulk* and *gate* at the given temperature. In order to neutralize the device, the total contact potential of the *gate*-insulator-*bulk* structure has to be overcome. The contact potential of a layered structure like this one depends only on the external materials.

An alternative view on the subject is by the means of the *work functions*[3]of the materials [10]: the energy bands bending is explained by the lineup of the *Fermi levels* of the materials at the interface. Therefore the *contact potential* of the junction

[1]The *Fermi potential* of extrinsic silicon is the difference between its *Fermi Level*(E_F) and the one of intrinsic silicon(E_i), expressed as a voltage:

$$\phi_F = \frac{E_F - E_i}{q} \tag{2.5}$$

Where q is the electron's charge. It is approximately $-0.56V$ for n-doped silicon and 0.56 for p-doped one.

[2]*Contact potential*: the potential that arises at the ends of a junction between two materials

[3]*Work function*: the energy required to move an electron from the *Fermi level* to the *vacuum level*, namely the energy required to remove an electron from the material.

can be put in relation with the materials *work functions*(W_1, W_2):

$$\phi_{12} = \frac{W_2 - W_1}{q} \tag{2.7}$$

The band diagrams in fig.2.4 represents this relation for *MOSFETs*.

Another voltage that needs to be neutralized is the one given by trapped charges inside the insulator. This can be accounted for using the concentration per unit area of this charges (Q_{ox}) and C_{ox}.
Finally, when a voltage is applied between *source* and *bulk* (V_{SB}), the so called *bulk effect* arises. The voltage between the contacts and the *bulk* reverse biases the *channel-bulk* junction widening the depletion region and therefore the exposed charge that needs to be neutralized. The complete expression for the *threshold voltage* is:

$$V_{th} = V_{th0} + \gamma\left(\sqrt{2\phi_F + V_{SB}} - \sqrt{2\phi_F}\right) \tag{2.8}$$

Where γ is the *bulk effect coefficient* that represents the device's sensitivity to the bulk effect:

$$\gamma = \frac{\sqrt{2\varepsilon_{Si} q N_A}}{C_{ox}} \tag{2.9}$$

Linear Region ($V_{GS} - V_{th} > 0 \wedge V_{DS} \leq V_{GS} - V_{th}$)

The characteristic of the *MOSFET* in this region is the following:

$$I_{DS|l} = \mu_n C_{ox}\left(\frac{W}{L}\right)\left[\left(V_{GS} - V_{th}\right)V_{DS} - \frac{V_{DS}^2}{2}\right] \tag{2.10}$$

As it can be noted from the equation, I_{DS} behaves as a resistor for small values of V_{DS}. Intuitively, the value of the resistance can be interpreted as the one of a conductor: growing linearly with its length and decreasing with its area, hence the $\left(\frac{W}{L}\right)$ term.

A non liner quadratic behavior is met for high V_{DS} values, this is attributed to the non uniformity of the *channel* caused by V_{DS} imposed field.

Saturation: Strong Inversion ($V_{GS} - V_{th} > 0 \wedge V_{DS} > V_{GS} - V_{th}$)

In the case in which V_{DS} assumes a large value compared to V_{th}, the carriers in the *channel* can no more sustain the *linear region* trend for I_{DS}. The current value almost saturates, therefore the *MOS* behaves as a voltage controlled current source.

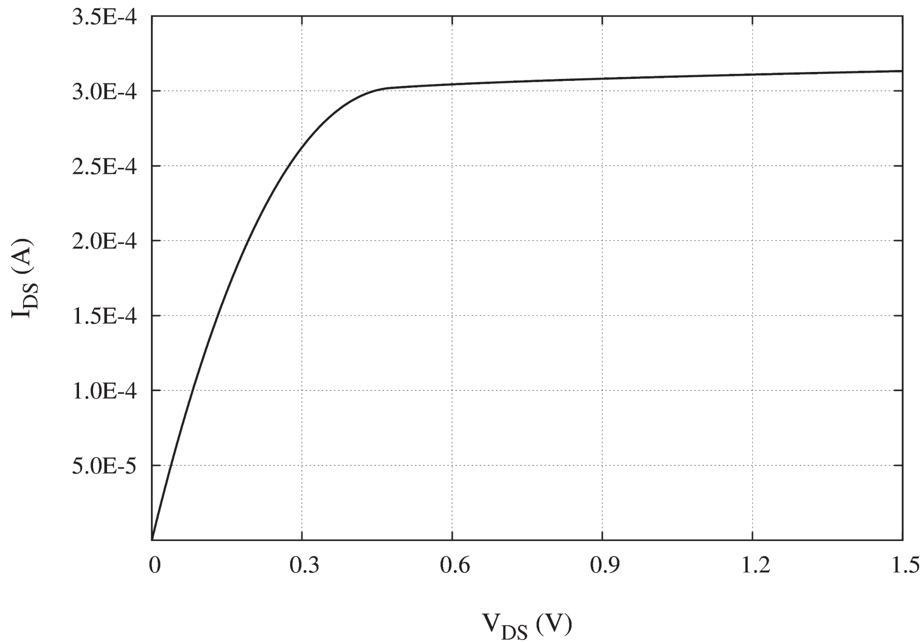


Figure 2.5: I_{DS} characteristics from liner region to saturation, taken from [8]

This characteristics is useful in analog circuits, but it must be considered in the design of circuits measures that keep the transistor in saturation.

The complete relation for I_{DS} in *strong inversion* is:

$$I_{DS|s} = \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)(V_{GS} - V_{th})^2(1 + \lambda V_{DS}) \tag{2.11}$$

The last part of the equation describes the *channel length modulation effect*: due to the *pinch-off* of the channel caused by external field of V_{DS}, the *effective length* of the channel decreases giving a weak dependence to I_{DS} on V_{DS}.

λ is the *channel length modulation effect parameter* and it relates the relative *channel* shortening with V_{DS}:

$$\lambda V_{DS} = \left(\frac{\Delta L}{L}\right) \tag{2.12}$$

This parameter defines the device's sensitivity to this effect. It must be noted that its bigger for devices with shorter *channels*, so it is not a technology defined parameter.

level of inversion	I_{DS}
strong	$$\frac{1}{2}\mu_n C_{ox}\left(\frac{W}{nL}\right)(V_{GS}-V_{th})^2 \quad (2.14)$$
moderate	$$2n\mu_n C_{ox}\left(\frac{W}{L}\right)\phi_T^2\left[ln\left(1+e^{\frac{V_{GS}-V_{th}}{2n\phi_T}}\right)\right]^2 \quad (2.15)$$
weak	$$2n\mu_n C_{ox}\left(\frac{W}{L}\right)\phi_T^2 e^{\frac{V_{GS}-V_{th}}{n\phi_T}} \quad (2.16)$$

Table 2.1: EKV models for I_{DS}

Saturation: Weak and Moderate Inversion ($V_{GS} \sim V_{th} \wedge V_{DS} > V_{GS}$)

Even for $V_{GS} < V_{th}$ there is a residual I_{DS} inside the channel before the *depletion* point, the strong inversion relation previously illustrated does not consider this fact as it depends only on the overvoltage. The correct trend for this operating region called *weak inversion* is exponential in relation to V_{GS}:

$$I_{DS|w} = \mu_n C_{ox}(n-1)\left(\frac{W}{L}\right)\phi_T^2 e^{\frac{V_{GS}-V_{th}}{n\phi_T}}\left(1-e^{-\frac{V_{DS}}{\phi_T}}\right) \quad (2.13)$$

In this condition, the charge inside the *channel* can not be no more calculated with the parallel plate capacitor approximation. That is because the *channel* region is not completely inverted, but it presents a depletion layer (with its associated capacitance) between the *channel* and the *bulk*. The $(n-1)$ term is the ratio between this capacitance and the one of the oxide layer; for its role in the equation, n is called *slope factor*.

The *bulk effect* contribution is provided by the explicit value of V_{th}.
Lastly, a third region of inversion is considered in between the *weak* and *strong* one: the *moderate inversion*. Both the two previous relations can't provide an adequate I_{DS} relation for the intermediate region of inversion and, moreover, the transition between the two regimes is not continuous. For these reasons the *moderate inversion relation* is obtained as an interpolation between the two characteristics and is presented in eq.(2.15).

A recap on the different I_{DS} is presented in tab.2.1, these are a set of *bulk* references models called *EKV model*, widely used in the field.

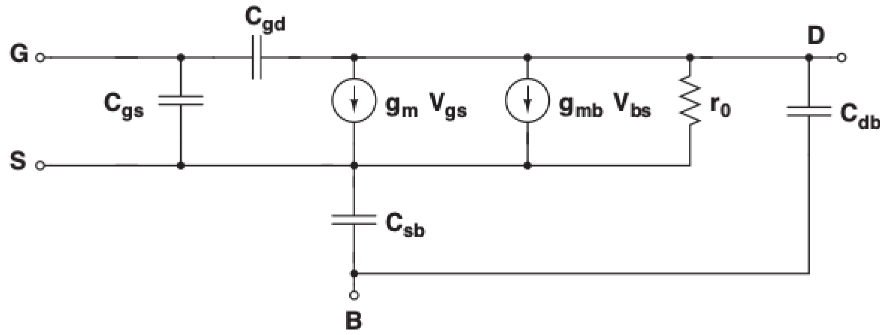


Figure 2.6: small signal equivalent of a *NMOS* transistor, taken from [8]

2.1.2 Small Signal Parameters

In order to study analytically electronic circuits, complex non-linear elements as transistors are usually linearized around the operation point. In this way, an equivalent circuit with ideal elements can be substituted to it, making the solution of the nodal equation simpler. This approach is called the *small signal equivalent* analysis of the circuit. For an electrical parameter $f(\vec{x})$ a *small signal parameter* $f_{sm,i}$ can be evaluated for every dependant variable x_i of its function at a certain operating point $\vec{x}^*$. The parameter can be obtained as the first order coefficient of the partial Taylor expansion of the function:

$$\begin{aligned} \Delta f(\vec{x}) &= \left(\frac{\partial f(\vec{x}^*)}{\partial x_i} \right) \Delta x_i \\ &= f_{sm,i}(\vec{x}^*) \Delta x_i \end{aligned} \tag{2.17}$$

It must be noted that this approximation holds as long as the signal is a small variation from the operating point and the signal must not move the transistor from one operating region to another. For example, a typical request is to keep all transistors in the circuit in *saturation* for all expected signals.

In fig.2.6 the schematic of the small signal model of a *NMOS* transistor, with the main capacitances affecting the circuit, is presented.

Gate Transconductance

This parameter represents the conductance of the equivalent ideal current source placed between D and S of a transistor in *saturation*, it's called *transconductance* because its value is voltage controlled by V_{GS}. It's trivial to use a *MOSFET* in this

configuration as the core gain element for a *transconductance amplifier*.
Its expression varies on the basis of inversion level.
For *strong inversion*, by definition:

$$
\begin{aligned}
g_{m|s} &= \frac{\partial I_{DS|s}}{\partial V_{GS}} \\
&= \mu_n C_{ox} \left(\frac{W}{L}\right) (V_{GS} - V_{th}) \\
&= \sqrt{2\mu_n C_{ox} \left(\frac{W}{L}\right) I_{DS}}
\end{aligned}
\tag{2.18}
$$

Here we can see its relation to $\left(\frac{W}{L}\right)$.
The last relation is particularly useful because the value of I_{DS} is typically chosen at design time in order to optimize the circuit power consumption versus area occupation: setting L, g_m grows linearly with the area and I_{DS}.
For the *weak inversion*, instead:

$$
\begin{aligned}
g_{m|w} &= \frac{\partial I_{DS|w}}{\partial V_{GS}} \\
&= \frac{I_{DS}}{n\phi_T}
\end{aligned}
\tag{2.19}
$$

As it can be noted g_m loses its dependence to the transistor aspect ratio in this regime, so it seams that the transistor can be made indefinitely small without compromising its g_m. However decreasing $\left(\frac{W}{L}\right)$ will move the transistor in *strong inversion*, regaining its dependence on the aspect ratio.

In order to quantify this situation it is necessary to define bounds for the inversion regions, for this reason the *boundary current* is introduced as the current value in which the strong and weak versions of g_m provide the same value:

$$
\frac{I_{DS|bound}}{n\phi_T} = \sqrt{2\mu_n C_{ox} \left(\frac{W}{L}\right) I_{DS|bound}} \rightarrow I_{DS|bound} = 2n\mu_n C_{ox} \left(\frac{W}{L}\right) \phi_T^2 \tag{2.20}
$$

As a reference, it's useful to use a technology dependant parameter as the *technology boundary current*:

$$
I_{DS|bound0} = 2n\mu_n C_{ox} \phi_T^2 \tag{2.21}
$$

Using the $I_{DS|bound}$ value, the *inversion coefficient* is introduced as the ratio between

the actual current and the *boundary current*:

$$I_C = \frac{I_{DS}}{2n\mu_n C_{ox}\left(\frac{W}{L}\right)\phi_T^2} \tag{2.22}$$

The inversion region is identified by its value:

$$\underset{\text{weak}}{I_C} \quad < \quad 0.1 \quad \leq \quad \underset{\text{moderate}}{I_C} \quad \leq \quad 10 \quad < \quad \underset{\text{strong}}{I_C}$$

As stated above, decreasing $\left(\frac{W}{L}\right)$ for the same I_{DS}, will move the transistor in *strong inversion*.
Finally g_m can be expressed in all regions using I_C:

$$g_m = \frac{I_{DS}}{n\phi_T}\frac{1}{\sqrt{I_C + \frac{1}{2}\sqrt{I_C} + 1}} \tag{2.23}$$

Bulk Transconductance

This parameter accounts for the *bulk effect* contribution to I_{DS}. It is calculated from the explicit dependence of V_{th} on V_{SB}:

$$\begin{aligned} g_{mb} &= \frac{\partial I_{DS}}{\partial V_{BS}} \\ &= \left(\frac{\partial I_{DS}}{\partial V_{th}}\right)\left(\frac{\partial V_{th}}{\partial V_{BS}}\right) \\ &= \mu_n C_{ox}\left(\frac{W}{L}\right)(V_{GS} - V_{th})\frac{\gamma}{2\sqrt{2\phi_F + V_{SB}}} \\ &= \eta g_{m|s} \end{aligned} \tag{2.24}$$

g_{mb} results proportional to g_m by a factor γ (typically $0.2 \div 0.3$).

Output Conductance

The output conductance is results from the *short channel effect* causing the dependence of I_{DS} from V_{DS}. The name is due to the lack of dependence of g_{ds} on

V_{GS}. Its explicit equation is:

$$\begin{aligned} g_{ds} &= \frac{\partial I_{DS}}{\partial V_{DS}} \\ &= \lambda \mu_n C_{ox} \left(\frac{W}{L}\right) (V_{GS} - V_{th})^2 \\ &= \lambda I_{DS} \end{aligned} \tag{2.25}$$

Due to its linear relation with I_{DS}, it's more straightforward to represent it as a resistor:

$$\begin{aligned} r_0 &= \frac{1}{g_{ds}} \\ &\simeq \frac{1}{\lambda I_{DS}} \end{aligned} \tag{2.26}$$

This is called *output resistance* of the *MOSFET*.

2.1.3 MOSFET Noise

Electronics noise stands for the statistical, unavoidable fluctuation of the signal. In this section the basic sources of noise in *MOSFET* are illustrated. Noise is studied with a statistical approach due to its random nature. Noise is expected to present a symmetrical distribution around the zero with a certain variance.

A useful quantity is the *noise power spectral density (PSD)* referred to the input of the circuit which is the distribution in frequency of the average power of a signal, for a generic signal $x(t)$:

$$\begin{aligned} S_{xx}(f) &= \lim_{T\to\infty} \frac{1}{T} |X_T(f)|^2 \\ &= \lim_{T\to\infty} \frac{1}{T} \left| \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) e^{-j2\pi f t} dt \right|^2 \end{aligned} \tag{2.27}$$

Where $X_T(f)$ is the truncated Fourier transform of the signal.
The fact that the noise is in this form and referred to the input, let us compute the influence of the circuit on the output noise density by the mean of a simple multiplication with its squared *transfer function*:

$$S_{yy}(f) = |H(f)|^2 S_{xx}(f) \tag{2.28}$$

If other independent noise sources are considered, their contributions to the *PSD*

are simply summed up.
For the sake of brevity the *PSD* will be further indicated with v_n^2 for voltage values and i_n^2 for current ones.

Thermal Noise

One of the primary noise sources in electronics is the variation in current density due to thermal agitation of carriers. For a resistor of value R the power distribution of this noise follows the Plank's law for black body radiation in one dimension [11]:

$$\begin{aligned} v_{nT}^2 &= \frac{4Rhf}{e^{\frac{hf}{k_BT}}-1} \\ &\simeq 4k_BTR \end{aligned} \tag{2.29}$$

Where the approximation follows from the fact that at room temperature the frequency at which the exponential argument becomes unitary is $\sim 6.25THz$, so it can be first order approximated in zero.

$$i_{nT}^2 = \frac{4k_BT}{R} \tag{2.30}$$

This relation is valid for any type of conductor with a certain resistance R, so in the case of the MOSFET we can use the inverse of the transconductance:

$$i_{nT}^2 = 4k_BT\gamma g_m \tag{2.31}$$

γ is called *inversion factor* and it takes into account the variation of the mobile charge in the *channel* in the different inversion levels:

$$\gamma = \frac{1}{2} + \frac{1}{6}\frac{I_C}{I_C+1} \tag{2.32}$$

The noise expected with this relation results too optimistic for short channel devices (under $1.7\mu m$ [12]): in this an *excess noise correction factor (α_w)* is introduced to take into account this effect:

$$i_{nT}^2 = 4k_BT\alpha_w\gamma g_m \tag{2.33}$$

This current noise can be seen as an output noise in all the applications which use the *MOSFET* as a transconductance amplifier; for this reason it may turn useful to

convert this quantity as voltage to be used as an input noise source:

$$v_{nT}^2 = 4k_B T \alpha_w \gamma \frac{1}{g_m} \tag{2.34}$$

Flicker Noise

This is another important noise source in *MOSFETs*, it's often called $\frac{1}{f}$ noise due to its *PSD*.
This contribution has been explained with two competing models: one ascribes this noise to interactions between carriers and phonons of the silicon lattice, the other one to fluctuation on the number of carriers due to trapping in the $Si - SiO_2$ interface.

The following relation unifies the two descriptions:

$$v_{nf}^2 = \frac{K_f(I_C, L)}{C_{ox} W L} \frac{1}{f^{\alpha_f}} \tag{2.35}$$

Here it can be noted that v_{nf}^2:

– varies with the inversion level thorough the $K_f(I_C, L)$ term

– it's not exactly $(\frac{1}{f})$ but it has correction term α_f

– it decreases with the transistor size

– is technology dependant

2.1.4 Mismatch in *MOSFETs*

Mismatch is the process that causes time-independent random variation in physical quantities of identically designed devices [13]. It must be pointed out that mismatch refers only to variations due to random process during fabrication stages of a single wafer: they are not batch to batch or wafer to wafer variations.

This can affect, concerning this work, analog signal processing and reference sources. The scaling of the devices in *CMOS* process amplifies these effects due to their dimension and reduction of signal swing, caused by the reduction of power supply.

It can be generally evaluated as a variation of a parameter $P(x, y)$ of two identical features occupying a certain area $A(x, y)$ in two different positions (x_1, y_1) and

(x_2, y_2):

$$\Delta P(x_{12}, y_{12}) = \frac{1}{A}\left[\int\int_{A(x_1,y_1)} P(x', y')dx'dy' - \int\int_{A(x_2,y_2)} P(x', y')dx'dy'\right] \quad (2.36)$$

It can be interpreted as spatial convolution of the effect of a spatial window and a random intrinsic variation of the parameter. Using the Fourier transform, it can be expressed as the product of functions of spatial frequencies:

$$\Delta\mathcal{P}(\omega_x, \omega_y) = \mathcal{G}(\omega_x, \omega_y)\mathcal{P}(\omega_x, \omega_y) \quad (2.37)$$

In this way $\mathcal{P}$ generates mismatch for certain frequencies that is then filtered by geometry dependence of the area $\mathcal{G}$.

There is a class of variations on P that are composed of many small independent high-frequency contributions than, once integrated in sufficiently large area, tend to delete each other. This category includes: the number of implanted or diffuse ions, substrate ions, local mobility fluctuations, oxide granularity and charges.

There is a second class with long range variations originated from deterministic effects dependent on position the on wafer. These are ultimately considered random because part of the deterministic information can be lost, like the position on the wafer after packaging. In this case the spatial frequency is proportional to $\left(\frac{1}{D}\right)$, where D is the wafer dimension.

For both classes the mismatch distribution is expected to be normal, so it can be expressed by the mean of its variance σ^2. Considering a rectangular window (the most common transistor shape) of area WL its variance results in:

$$\sigma^2(\Delta P) = \frac{A_p}{WL} + S_p^2 D_x^2 \quad (2.38)$$

The first term accounts for high frequency variation by the mean of the coefficient A_p, thus it decreases enlarging the area. The second one with its coefficient S_p, accounts for the low frequency terms it therefore depends on the wafer dimensions.

In terms of *MOSFET* parameters we are interested in its characteristic I_{DS}: it depends on three process dependant factors in all regions:

- V_{th0}: the threshold voltage in absence of bulk effect. It varies with variations on: oxide charges Q_{ox}, dielectric granularity (C_{ox}) and depletion charges depending on the number of dopants N and affecting ϕ_G and ϕ_B. These are all uncorrelated and follows (2.38).

- γ the bulk effect coefficient varying with N and C_{ox}. In the same way as the

previous it follows (2.38).

– β the *current factor* defined for convenience as:

$$\beta = \mu C_{ox} \left(\frac{W}{L} \right) \tag{2.39}$$

Which variations on mu and C_{ox} follows (2.38), while $\sigma_W^2 \propto \frac{1}{W}$ and $\sigma_L^2 \propto \frac{1}{L}$. Combining them to explicit the WL term shows that $\left(\frac{W}{L}\right)$ contribution is of the second order in W and L so it can be neglected.

In summary, at the first order, the mismatch contributions from *MOSFETs* follow the equation (2.38). The entity of theses contributions is represented by the related coefficient A_p and S_p that are technology dependent and extracted empirically from fit of experimental data.
In order to mitigate these effects, the layout of sensitive circuit portions can be designed with some precautions like subdividing single transistors in smaller ones and spreading them in a larger region in order to increase their effective area.

2.2 *CMOS* Fabrication Process and Design

Complementary MOS (*CMOS*) is a technology for constructing integrated circuits. The name comes form the possibility offered by this technology to manufacture *PMOS* and *NMOS* device on the same substrate.

The advantages of *CMOS* are evident in digital circuits since it enables to produce logic gates that don't dissipate current while static, but only during transients. In terms of analog circuits this technology enables the implementation of high performance amplifiers as they present small capacitance for high bandwidth and it makes possible to substitute most of the passive elements with transistor in various configurations.

The reasons of the wide adoption of this technologies are the versatility of its manufacturing processes and the continuous downscaling of the feature size throughout the years that enables integrating more features in the same chip area.

In this section first a brief description of the *CMOS* technology is presented in 2.2.1, than in 2.2.2 the typical manufacturing process for technology up to the 130nm node is detailed with a brief view of target node specifics. In the next subsection, 2.3.1, various scaling methodologies and modern techniques are presented. Finally in 2.2.3 a typical design flow for *CMOS* analog integrated circuit is illustrated.

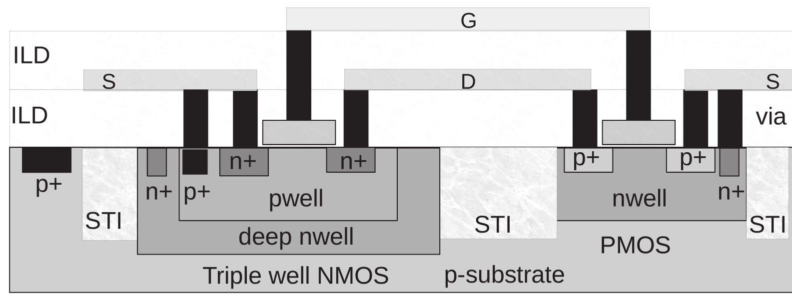


Figure 2.7: *CMOS* devices cross section, taken from [8]

2.2.1 *CMOS* technologies

CMOS circuits are organized in layered structure stacked upon the same silicon substrate, a cross section representation of *NMOS* and *CMOS* transistors in *CMOS* technology is presented in fig.2.7. The first layer from the substrate is called *active area*, on this layer transistors and other devices are realized.

The substrate is typically p-type since it's more cost effective and the low mobility of the holes reduces the noise coming from other part of the circuit (*substrate noise*).

The substrate behaves as a common *bulk* for *NMOS* devices; in order to have a *bulk* terminal for *PMOS* devices a *nwell* is created by doping n^+ the *PMOS* area in order to have an effective n-type region.

In order to create exclusive *bulk* contacts for *NMOS* s, p-type regions must be insulated from the substrate. *Triple-well* devices can be made for this purpose by forming a region of p^+ doped silicon inside a n^+ well, called *deep nwells*. In this way an effective p doped region is created insulated from the substrate. Insulation is provided revers biasing the junction between the wells, so *nwells* require contacts (usually tide to high voltage).
Having different doped regions close together creates parasitic *BJT*s formed by series of junctions: a *npn* between D or S of a *NMOS* and substrate-*nwell*; and a *pnp* in the *PMOS*'s electrodes-*nwell*-substrate junctions. These two *bjt*s forms two common emitter amplifiers connected in a positive feed-back loop through the substrate resistance; this generates an instability that is called *latch-up*. In order to prevent this phenomena the two *MOS* needs to be placed at a minimum distance.

Many solutions are employed to mitigate *latch-up* and substrate noise like using different substrate doping profiles in order to increase substrate resistivity, or to insulate analog parts or single devices by using layers and tranches of insulator (*shallow tranches* and *silicon on insulator* isolations).

Regarding connectivity, nearby electrodes can be connected together by simply

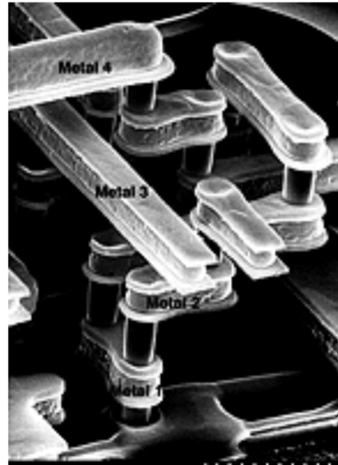


Figure 2.8: Electron microscope photo of four metal layers, taken from *Lecture 5 on Digital Integrated Circuits, J. M. Rabaey , A. Vladimirescu, Berkeley University, 2002*

making common diffusions; on the other hand different kind of connections must be made by metallization (usually in Al or Cu) in upper dielectric layers. The dielectric material is usually SiO_2 and thees layers are called *Inter Level Dielectric (ILD)*, they serve as insulators for the metallic line which are embedded in them. Upper metallization presents wider and thicker lines in order to have less linear resistivity, this is done to reduce the total resistance since these tend to be the longest connections of a circuit.
Vertical connections between subsequent layers are called *Vertical Interconnection Access (VIA)*, they are metal connections that connect the first metal layer with the *active area* as well.

2.2.2 Manufacturing Process

The conventional *CMOS* manufacturing process is done in different steps in order to form the horizontal layout of the device and the vertical structure of the interconnections layers.

In general to create a certain pattern the process consists in:

1. A layer of SiO_2 is formed by oxidation of the top layer, using oxygen in a 1000°C furnace.
2. A layer of *photoresist* is formed. It is a polymeric material that can be made vulnerable to chemical attacks by illumination.

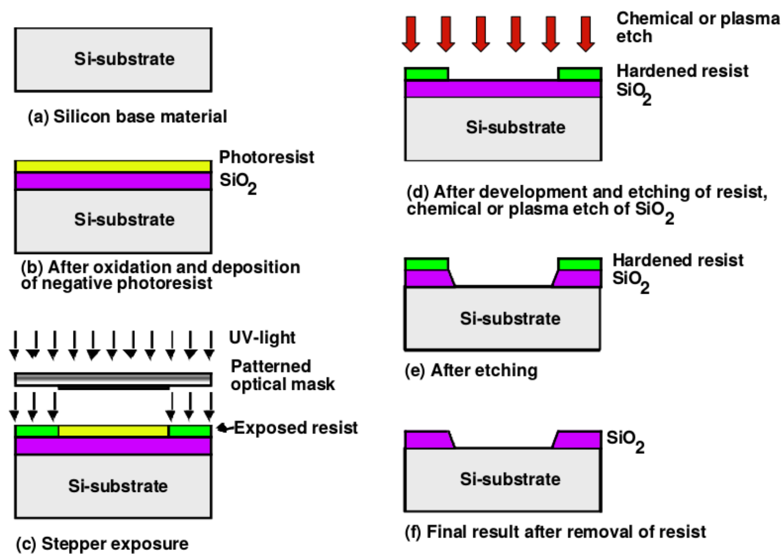


Figure 2.9: Scheme of device patterning, taken from *Lecture 5 on Digital Integrated Circuits, J. M. Rabaey , A. Vladimirescu, Berkeley University, 2002*

3. The *photoresist* is exposed to the light through a binary mask and a shrinking lens, where the corresponding pattern is the desired one.

4. The exposed *photoresist* and the SiO_2 underneath it are removed using a chemical attack.

5. The exposed zones are processed by various techniques depending on the type of the wonted material.

6. The residual *photoresist* is removed with a new chemical attack.

A scheme of this patterning procedure is presented in fig.2.9.
In this way the different regions of the active area can be made in precise order:

1. deposition of eventual insulating tranches by plasma etching of the substrate p-doped material, filling with the tranches insulator and planarization with *Chemical Mechanical Polishing (CMP)*.

2. *nwell* doping by diffusion or ion implantation.

3. *pwell* doping by diffusion or ion implantation.

4. *gate* material (*polysilicon*) deposition and etching.

5. implants of *drain* and *source* implants, for both p^+ and n^+ doping.

6. deposition of *ILD* and patterning for *VIAs*.

7. deposition of the metal layer and *CMP*.

The 5 to 6 steps are repeated for every metallization layer.

2.2.3 Design Cycle

The complexity of *CMOS* design is mainly due to two factors: high cost of manufacture of the first prototype (mainly the cost of masks set for lithography) and the difficulty to probe specific circuit node in the prototype (because of the physical dimension of them and due to chip packaging).

To solve the second problem, testing structures with dedicated pins can be planned for preliminary test chips. But the number of pins is constrained by chip area and number of pins (proportional to the chip perimeter).

So a good solution for both problems is to have robust simulation mechanisms that allow to design working prototypes on the first attempt. These tools are called *Electronic Design Automation (EDA)*, we will focus on the analog side of the design using these tools.
Firstly a circuit simulator is necessary: this software will solve the nodal equation of the circuit providing a numerical simulation of the circuit operation. The single devices are simulated using manufacturer provided model libraries, this models can be physics based or, as in modern nodes, partially empirical. The designer has to draw the schematic of the circuit defining the eventual parameters, he also can add ideal components for testing. The type of analysis ranges from DC, AC, transient, noise and many other.

There's also the possibility to take into account a variabilities on device characteristics, this is done mainly in two ways: corners and montecarlo. The first one considers the edge condition in terms of speed of operation for both *NMOS* and *PMOS* called corners, and thus simulations can be made on the desired corner or as a combination of them. The montecarlo models, instead, rely on statistical distributions on devices parameters, both for process and mismatch variations. The analysis is therefore repeated with randomized values picked from those distributions, and the result itself is as distribution of a wanted parameter. It is clear that this approach becomes really time and computing consuming: in order to have statistically complete sets the number of repeated runs will grow with the combination on the

number of variables. It's frequently more convenient to run montecarlo on smaller subsets of the circuit sensible to these variations.

The next step consists in actually drawing the layout of the circuit with a specif CAD (Computer Assisted Design), the output of this step are various bidimensional maps which are used by the manufacturer to create the actual masks for every layer of the desired process technique.

As explained throughout this section, the manufacturing process has some strict rules that need to be observed to make the techniques applicable; this applies even more on newer technologies as explained in section 2.3. In order to help the designer in this operation a software called *design rules checker* will monitor if the layout follows those rules.

The layout needs to be equivalent to the schematic representation previously built, so another software checks the equivalence of the devices and their connectivity.

Finally another software extracts possible parasitics elements from the design like capacitance between nearby lines or the total resistance of any single line.

Having said that, a typical process aware design for analog circuits can be structured as:

1. preliminary circuit design using small signals equivalent models, due to their analytical simplicity.

2. design correct operation check and optimization (DC levels, AC transfer functions, power consumption, transient timing) using schematic simulations.

3. design optimization for noise and variation in process, supply voltages and temperature using corners or montecarlo analysis.

4. circuit layout.

5. design rules check.

6. connectivity checks.

7. extraction of parasitics.

8. schematic simulation with parasitics added in.

From every step, if desired performances and constrains are not met, the designer may be forced to turn back to a previous one if design changes are needed.

2.3 Scaling to the 28nm Technology node

This subsection will discuss the changes in *MOSFET* structure adopted by the manufacturers in order to maintain the yearly trend in scaling factor while keeping the performances in line with the previous iterations.
The 28nm target technology of this study was designed to overcome many shortcomings of the Denard Scaling, as the increase of the tunnel effect induced gate current through the thin insulator. In 2.3.1 this and other scaling problem will be discussed along with the characteristics of usual *CMOS* scaling.
The techniques used in modern process node include the usage of new materials for the insulation layer as *High-k dielectrics(HK)*, as well as various techniques to enhance carriers mobility via material straining by the insertion of *SiGe* in the electrodes area. A TEM micrograph of of a *PMOS* transistor with these features is reported in fig.2.10. These two features are respectively discussed in 2.3.2 and 2.3.3 .

Since the light's wavelength used in common lithography is comparable to the features size, resolution enhancement techniques are necessary, some of these are discussed in 2.3.4.

The selected technology node features an HK layer on top of a 2nm thick SiO_2 layer, in particular it is an hafnium oxide based (HfO_2) material. The gate material is TiN for the *NMOS* and $TiAlN$ for the *PMOS*. The *PMOS drain* and *source* electrodes material is *SiGe*. A TEM micrograph of a *PMOS* device with these features is presented in fig.2.10 This technology also requires a regular layout for transistor's gate which imposes a constant separation pitch within all parallel gates of at least 110nm. These constrains come from the adoption of *phase shifting masks* for the lithography. The technologies comes with up to eleven metallization layers for interconnection, the last two of them reserved to non-dynamic signals.

2.3.1 Scaling

As said before, scaling is a fundamental factor in electronics industries.
It enables to built more circuits in the same area, to make them more power efficient and to decrease their cost as more of the replicas of the same chip can be made in the same wafer area.

An ideal scaling would be achieved by scaling both the horizontal and vertical dimension of the chip by certain scaling factor $s > 1$(*Dennard scaling*). Although not every aspect of the transistor scales linearly with the geometry, for this reason scaling must be moved from the ideal one.

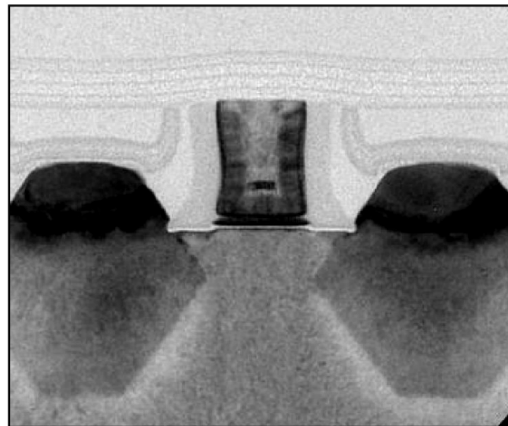


Figure 2.10: TEM micrograph of a *PMOS* with HKMG and *e-SiGe*, taken from [14] *©2007 IEEE*.

A scaling methodology is the *constant filed* one: every geometrical feature is shrunk by $(\frac{1}{s})$ and at the same time operating voltages are reduced to compensate the increasing of the electric field in the material. The first consequence is the growth of the insulator capacitance per unit area C_{ox} as due to reduction of t_{ox} as in eq.(2.1):

$$C_{ox} \rightarrow sC_{ox}$$

In this way all C_{ox} dependent parameters will be revers affected by the scaling, like the *technology boundary current* would increase according to eq.(2.21): in this way a more intense current will be required in order to keep a transistor with the same aspect ratio in *saturation*.

Furthermore, all the depletion regions inside the material don't scale with the geometries, so, in order to shrink them, the dopants concentrations must be increased by s degrading the carrier mobility.

An exception is represented by g_m in strong inversion in which the contribution on the voltage and t_{ox} scaling are compensated.

The limit of the *constant filed scaling* is dictated by V_{th}: since the *subthreshold slope* of I_{DS} is not technology dependant, the leakage current of the switched-off transistor (I_{off}) will be bigger and will cause grater static power consumption.

The opposite method is *constant voltage scaling*, that consists in leaving unchanged the voltages and let the electrical filed increase. The downside of this approach is that it will lead the transistor in breakdown. So a proper solution can be a compromise of the two methods.

Until the 130nm node the devices were scaled by a $\times 0.7$ factor every two years($s \sim$

1.4), but reached the 90nm node the thin layer of few atoms SiO_2 exhibits an excessive leakage current[14]. The origin of this current is in the *tunnel effect* that let the electrons pass through the dielectric potential barrier with a non negligible probability.

In order to counteract this effect new insulator materials were investigated: increasing their relative dielectric constant k the potential barrier is increased, thus the tunnel probability diminishes.
For this reason t_{ox} is no more a technology defining feature, but has been replaced with the *Equivalent Oxide Thickness (EOT)* defined as:

$$EOT = t_{hk}\left(\frac{k_{ox}}{k_{hk}}\right) \tag{2.40}$$

Usage of new materials makes possible to obtain the wanted EOT and thus the $\times 0.7$ shrinking factor, while reducing by at least $\times 0.25$ the tunnel effect leakage current.

Other problems of extremely scaled *MOSFETs* are the mobility degradation (that poses limits to L_{eff}) and the increase of parasitics capacitance due to transistor proximity. These effects are mitigated with strained channel material and increased gate pitch [15].

2.3.2 High-k dielectric with Metal Gate

As already mentioned in the section's introduction the *tunnel effect* problem has been resolved adopting HfO_2 as the insulator HK material for technologies with a L_{min} of 45nm or less. In this way the I_{off} was reduced by a $\times 25$ factor for the *NMOS* and a $\times 10^3$ factor for the *PMOS*, furthermore they exhibit a reduced V_{th} rolloff [14] due to large W values.

The first problem on the usage of HfO_2 consisted in founding a proper process technique to deposit a thin well controlled layer of it, this result is obtained by the standard SiO_2 formation by oxidation which is then removed and replaced with the HK insulator via *Atomic Layer Deposition (ALD)* .

Another problem comes with the *polygate*: the thin insulator layer lets the electric filed inside the gate forming a depletion layer that prevents the V_{th} control by doping. For this reason even for the *gate* new metallic material were investigated due to their screening effect from electrical fields.
The choice of the metal depends on other factors: as discussed in 2.1.1 the proper value of V_{th} can be tuned acting on the gate material *work function*, since metal materials are not silicon based, doping is no more a solution. In order to solve the problem, two different metals need to be used for *PMOS* and *NMOS* device in order

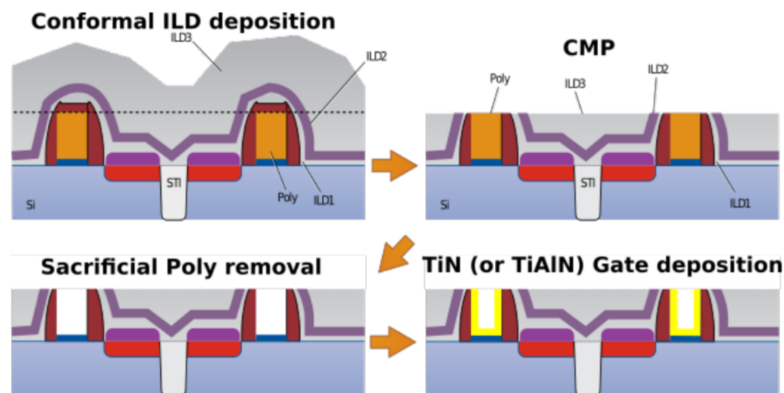


Figure 2.11: Schematic representation of *gate-last deposition*

to obtain a proper *work function*: for the conduction band (4.4eV) and the valence band (5.15eV) respectively.

When a metal with the proper working function is put in contact with the silicon on the channel a new issue arises: the Fermi level on the junction is pinned to an energy close to the middle of the band gap irregardless of the metal material (*Fermi pinning*). This phenomena is caused by the formation of spurious S-O bonds in the interface layer, in order to have a good interface purity rapid anneal at 1050°C is needed [16].

The chosen metals for the gates are TiN for the $NMOS$ and $TiAlN$ for the $PMOS$, as for the insulator the new process requires a first deposition of *polysilicon* that will be replaced with the proper metal (*gate-last* deposition, schematized in fig.2.11).

In summary, a complete process for a *High-K dielectric Metal Gate transistor (HKMG)* is performed by:

1. formation of SiO_2 layer by oxidation of the substrate.

2. the SiO_2 is removed with a chemical attack.

3. HfO_2 layer is formed by ALD.

4. deposition and patterning of the *polysilicon*.

5. opening of the *polygate* by planarization with CMP and polysilicon removal.

6. deposition of the $PMOS$ metal gate.

7. patterning of the $NMOS$.

8. *NMOS* metal deposition.

9. fill of the metal (Al) contacts.

In order to have a good planarization for an efficient *polygate* opening, the layout must be sufficiently homogeneous on the horizontal material distribution of this first layer (mainly gates and VIAs).

2.3.3 Channel Strain

Mechanical strain on silicon is adopted to enhance the mobility of carriers inside the material. This can be done for two reasons: to counteract the mobility reduction of the scaling (increased doping) and to equalize the difference in mobility of holes and electrons, discussed in 2.1.

There are two different kinds of strain: biaxial, which increases the mobility on the entire plane; uniaxial for an increment in a specified direction (in case of *MOSFET* the direction of L is the crucial one).

In the case of HK insulated device the typical biaxial solutions like inserting a buried under the channel compressed material (*SiGe*) can cause an increase on the gate leakage current due the presence of defects on the insulator interface [17][18].

A technique for uniaxial strain is the so called *Embedded SiGe (e-SiGe)*[19]: the strain is caused by *SiGe* made D and S electrodes: the D and S material is etched and replaced with *SiGe* growth epitaxially, that strains uniaxially the material of the channel between these two regions. With this technique an increase of 45% is expected for peak μ_p.

In order to strain the *NMOS* channel a silicon nitride cap can be added on top of the *gate*, in this way the two type of *MOSFETs* can be engineered separately.

To complete the process order of the previous subsection, the described process for the *e-SiGe* must be added after the *polygate* deposition [14].

2.3.4 Resolution Enhancement Techniques (RET)

The scaling on modern node has pushed the minimum feature of the design beyond the wavelength (λ) used for the photolithography that is employed to manufacture them. With these dimensions, the light passing through the aperture of the mask will produce interference effects: small aperture diffraction and interference between adjacent apertures. The interference issues poses a limit to the minimum

feature produced using standard lithography, in fact the maximum resolution obtainable with coherent light expressed in spatial frequency (ν_c) is:

$$\nu_c = \frac{NA}{\lambda} \tag{2.41}$$

Where $NA = n\sin(\theta)$ is the *numerical aperture* and θ is the angle formed by the optical axis and the lens radius. From these two equations the two main methods used to increase the resolution will be derived: increase NA or decrease λ.

The first method has been pursued through the year by increasing the lens radius, limited by the biggest reliable lens that can be built or the shortest distance between lens and focal plane; or by using other mediums to replace air. Using these methods as led to a natural progression to wavelengths of 193nm (in place of 365nm in the eighties); the limit on the decrement of λ is posed by the opacity in shorter wavelengths of the quartz used for the lenses.

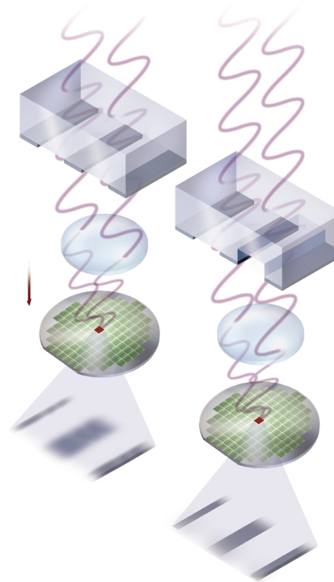


Figure 2.12: Principle of PSM: left standard lithography, right with PSM. Taken from [20] ©*2003 IEEE*

To overcome this limit the industry uses *Resolution Enhancement Techniques (RET)* [20]: this methods use constructive and destructive interference in order to move the resolution beyond the interference limit.

The first technique is called *optical proximity correction (OPC)*: it consists in modifying the aperture on the masks in order to compensate diffraction effect in corners and other sensitive shapes. The designer of the layout can be completely anaware of these corrections, therefore this method is applicable while leaving the design unchanged.

Off Axis Illumination (OAI) increases resolution by using holographic windows on the mask which irradiate light with different angles aimed to counteract diffracted rays with destructive interference.

The technique used in the technology selected for this work adopts the *Phase Shifting Mask (PSM)* [21]: masks which couple every aperture with another one which shifts the phase of the light in order to use interference effect to increase

or decrease the exposition of selected wafer areas. This method was initially used to prevent diffractive issues and it grants a theoretical resolution increment of $\times 2$. But, if the design is regular, adjacent apertures of design can be used to correct their respective interference, in the modern processes masks with etched quartz windows of different thickness are used. A representation of PSM principle is illustrated in fig.2.12.

Both OAI and PSM require repetitive patterns in the design's layout in order to operate properly, but they can work in conjunction with OPC. Using both method has led to an increase on the resolution up to $\frac{1}{8}$ of λ ($\sim 24nm$).

2.3.5 Impact of Scaling on Analog Circuits

Scaling operated with digital applications in mind, resulting in faster, more compact and less power hungry gates. It's inevitable to expect some shortcoming in terms of analog devices.
One aspect in which analog circuits benefit in the same way as the digital ones is in speed of operation due to reduced gate capacitance.

The first consequence of *CMOS* scaling is the reduction of power supply: throughout the years it was reduced from 5V to 1V. This makes it more difficult to keep all the transistors in saturation due to the reduced voltage headroom. Another consequence is that with the same architecture, analog amplifiers (discussed in 2.4.2), tend to have worst gain with smaller V_{DS} [22]. To counteract this effect more current can be used per branches, thus increasing the overall power consumption.

Another problem was caused but the increasing of the leakage current with the reduction of t_{ox}. This is no more an issue with the adoption of HK dielectrics.

In order to make the openings of the gates correctly with CMP, the layers must be really homogeneous: this introduces more design rules like minimum and maximum area of a certain material feature.
The adoption of new techniques has made the lithography less precise, thus it increases the mismatch. Moreover PSM requires high regularity for the design: the layout must be composed of regular zones of same sized and same spaced gates, in the edge of the zone dummy transistor must be made in order to maintain this trend, as illustrated in fig.2.13 In any case the area usage of analog circuit in these technologies does not scale with the same factor of the minimum features.
The interconnections have not scaled with the same factor of the transistors, this makes it more difficult to integrate design with a big number of connections. For the

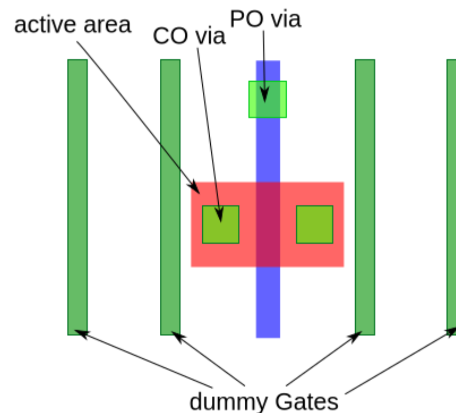


Figure 2.13: Simplified representation of a transistor layout with two dummy gates for each side.

same reason the metal density is increased along with the parasitic capacitances.

For modern nodes the same model for the 130nm is used, results are derived by scaling EOT, L_{eff} and V_{th} and mobility [23], but this approach may be not suitable for devices so different from the standard one.

The small dimension of VIAs makes their filling less reliable requiring some redundancy on their number.

These issues make the first schematic simulation a weaker tool to predict the final circuit behavior, in this way the normal design cycle will be more prone to backtrack because of unmet performance due to parasitics or mismatch; or for layout impossible to make because of space concerns or design rules.

2.4 Basic *CMOS* Analog structures

This section introduces some essential analog structures that can act as primary building blocks for most *CMOS* analog circuits or, at least, the one designed for this work (presented in chapter 3).

An essential device analog circuits is a voltage amplifier, for this reason in 2.4.2 single ended amplification stages are presented. In section 2.4.3 differential stages, circuit sensitive to signals relative variation, are discussed. Section 2.4.1 presents some useful building blocks that replace passive elements with *CMOS* configurations or work as real implementation of ideal circuit elements.

All the analysis are carried out using small signal equivalent circuit to study signal variation irregardless of DC levels. In this approximation all static nodes in

terms of voltage are replaced by a small signal ground, and all the transistor with their small signal equivalent presented in 2.1.2.

2.4.1 Buffers and Circuit Biasing

Source follower (fig.2.14a)

This circuits is a voltage buffer that is needed in order to mask sensible high impedance nodes from external low impedance nodes. The circuit is so called because the voltage as its output is equal at the one imposed at its gate minus V_{GS} In order to not be perturbed it must an output impedance as low as possible:

$$\begin{aligned} Z_{Osf} &= \frac{r_{01}}{1+(g_{m1}+g_{mb1})r_{01}} \\ &\simeq \frac{1}{g_{m1}+g_{mb1}} \end{aligned} \tag{2.42}$$

Increasing g_{m1} (acting on $\left(\frac{W}{L}\right)$ or I_{DS}) improves the driving strength of the source follower without disturbing its output. In fact the current mirror transistor M_2 forms with M_1 a feedback configuration that locks the stage gain:

$$A_v = \frac{g_{m1}}{g_{m1}+g_{mb1}} \tag{2.43}$$

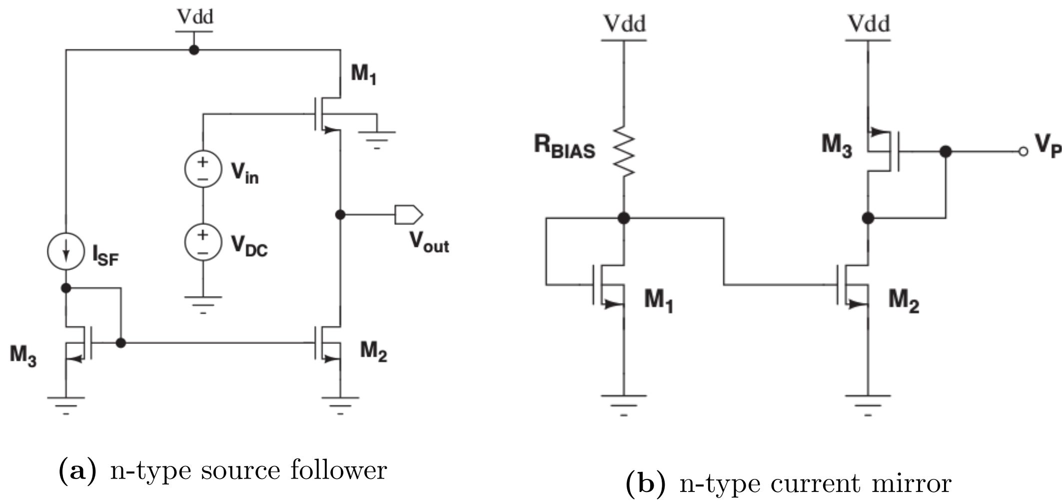
Having access to *triple well* devices move the gain to 1, creating a good buffer.

Current Mirrors (fig.2.14b)

The current mirror is a configuration that aims to provide a real current source terminal at the drain of transistor (M_2).

Transistor M_1 is diode-connected (its *gate* and *drain* terminal are short circuited), at its *drain* a certain current (I_1) is injected by a reference current source. In this condition the DC voltage at the drain of M_1 the one that let flow the imposed current and can be computed using (2.10) or (2.11) depending in its operating region. Although the diode connection ensures that $V_{DS} > V_{DS|sat}$ that creates an equilibrium point in strong inversion. This voltage (V^*) is the one that produces, for that particular transistor, an I_{DS} in strong inversion equal to the imposed one.

So by the fact that in DC a transistor *gate* absorbs a very small current, this voltage can be utilized to drive another identical one without disturbing the system. Since, with a fixed V_{th}, I_{DS} depends linearly on $\left(\frac{W}{L}\right)$, transistor of different aspect



(a) n-type source follower

(b) n-type current mirror

Figure 2.14: Taken from [8]

ratio con be used to obtain currents multiple of the reference one:

$$\frac{\left(\frac{W_1}{L_1}\right)}{\left(\frac{W_2}{L_2}\right)} = \frac{I_1}{I_2} \tag{2.44}$$

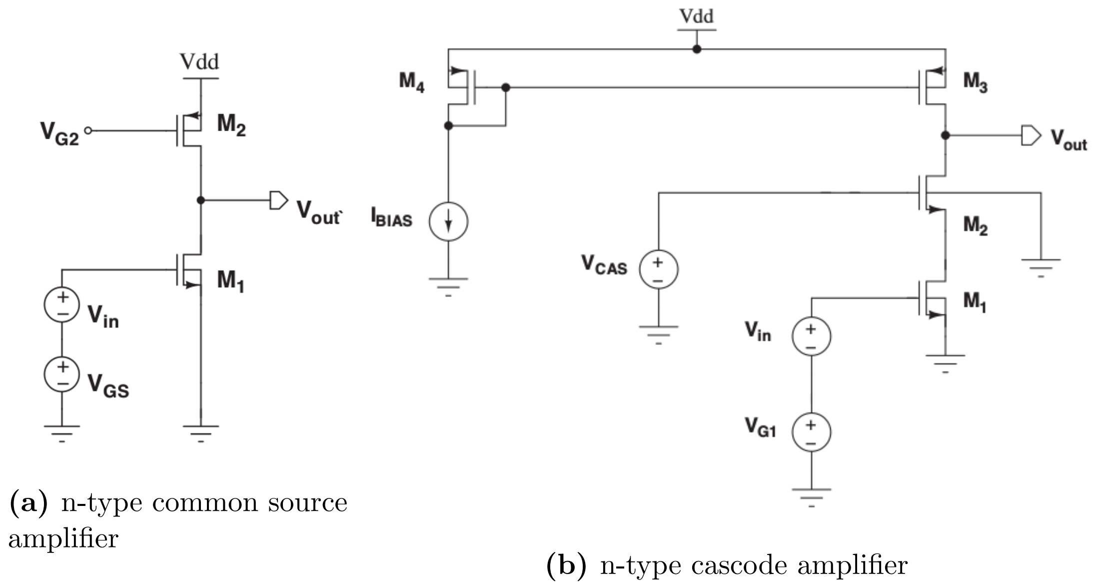
The only request to the circuit that needs to be driven by M_2 is to not push M_2 out of saturation.

Regarding the ideal current source, it can be obtained using a good voltage supply (with low output impedance) in series with a resistor (which can be connected off-chip).
Another aspect to consider is that one reference branch can be shared for more than one current mirrors, in the case of large integrated circuits the different interconnection lines resistance coupled with the small gate leakage currents can leads to unequal bias voltages.

2.4.2 Single Ended Amplification Stages

Single ended amplification stages are necessary in order to build good voltage amplifiers: devices with high gain, high bandwidth (BW), high input impedance and small output impedance.

This section discuss the first two points while taking into account important



(a) n-type common source amplifier

(b) n-type cascode amplifier

Figure 2.15: Taken from [8]

factors for pixels detectors: small power consumption, low-noise, small variability and small area occupation.

Obtaining a good input impedance is trivial in *CMOS* technology since, as explained later, the input node will be the *gate* of a transistor: its insulator operates as large impedance. Regarding the last point, source followers can be used for that purpose.

Common source amplifier with active load (fig.2.15a)

The common source amplifier is the simplest *CMOS* amplifier. As the name suggests its gain unit is a *MOSFET* with the source tied to ground.

Its operation mechanism is in principle a modulation of the current which flows through its channel's branch by V_{GS} via g_m. This current variation needs than to be converted to a voltage by a load.

Using passive resistors is not practical because integrating them is not convenient due to their big area occupancy. For this reasons the load is usually implemented with a transistor, in the case of *NMOS* is a *PMOS* and vice versa: in this way both the sources are tied to reference voltages.
The load resistor is in *current mirror configuration* (described in 2.4.1) so it can act both as the current source and as a load. The current source is essential because it

keeps the transistor in saturation and supplies the current which will be modulated by the input transistor.

For the small signal equivalent V_{DD} is a ground, so the parallel of their *output resistances* is the actual load. Thus the gain of the amplifiers is:

$$A_v = -g_{m|in} \left(r_{0|in} // r_{0|L} \right) \tag{2.45}$$

Where with $_{in}$ and $_L$ are indicated the input and load transistors respectively. As can be noted this is an inverting amplifier.

In order to make the amplifier operate properly both transistor need to be in saturation: the input one modulates linearly the current supplied by the load one only in in this condition; and the load one act as a current source in the same way.

From a designer point of view, different variables can be adjusted to build the desired amplifier:

- increasing $\left(\frac{W}{L}\right)_{in}$ for a fixed L increases both the gain and C_{gd}, thus limiting BW: $\left(\frac{W}{L}\right)_{in}$ constitutes a gain-BW trade-off.
- $\left(\frac{W}{L}\right)_l$ can be made smaller in order to increase its load resistance.
- increasing the channel area reduces the flicker noise contribution but as noted before reduces the BW: $(WL)_{in}$ constitutes a noise-BW trade-off.
- for the same reason flicker noise poses limits to the minimum area of both transistors
- thermal noise decreases increasing g_m, so the desired noise level poses a lower limit to $\left(\frac{W}{L}\right)_{in}$ and $\left(\frac{W}{L}\right)_l$.
- process variation is less sever in bigger devices: specific matching and repeatability necessities will pose a lower limit to $(WL)_{in}$.
- g_m is linearly dependent on I_{DS} in weak inversion: the choice of branch current is a gain-power consumption trade-off.
- from a DC point of view M_{in} and M_L forms a voltage partitioner: their relative sizing defines the output DC level.
- obviously making big transistors takes up large chip area, so any of the previous considerations that suggests big transistor creates a trade-of with chip area.

This arguments holds for any kind of *CMOS* amplifier: this topology represents the base for the others since any gain stage need a transconductance gain stage, a load to convert it in a voltage drop and a biasing current.

However this simple design presents two main issues: small gain (∼50) and a big output impedance. Having large open loop gain is a request for good feed-back systems. The impedance problem of the common source amplifier can be understand considering the case in which a small load i connected at its output: the impedance of an external load will be in parallel with the internals one creating an overall small load thus decreasing the circuits gain. More complex architectures where developed to remove this shortcomings.

Cascode amplifier (fig.2.15b)

This topology was introduced to improve the gain of the common source amplifier by adding an active load connected between its drain and the output of the amplifier. The load is realized with a transistor of the same type of the input one, the bias voltage (V_{bias}) applied at its gate is such that the selected I_{DS} keeps the transistor ins saturation. In this way, if *triple-well* devices are not viable, the *bulk* is not shorted with the *source*. The output impedance of this configuration is given by:

$$r_{out} = r_{01} + r_{02} + (g_{m2} + g_{mb2})r_{01}r_{02} \tag{2.46}$$

In *triple-well* is viable, the g_{mb} term is neglected. The transconductance of the cascode is given by:

$$\begin{aligned} G_m &= \frac{I_{out}}{V_{in}} \\ &= g_{m1}\frac{r_{01} + (g_{m2} + g_{mb2})r_{01}r_{02}}{r_{01} + r_{02} + (g_{m2} + g_{mb2})r_{01}r_{02}} \end{aligned} \tag{2.47}$$

This transconductance is lower than the one of the single transistor, but generally $r_{01} + r_{02} << (g_{m2} + g_{mb2})r_{01}r_{02}$ (moreover in absence of *bulk* effect), so G_m is close to g_{m1}.
By definition the gain of the full amplifier is:

$$A_v = G_m(r_{out}//r_{03}) \tag{2.48}$$

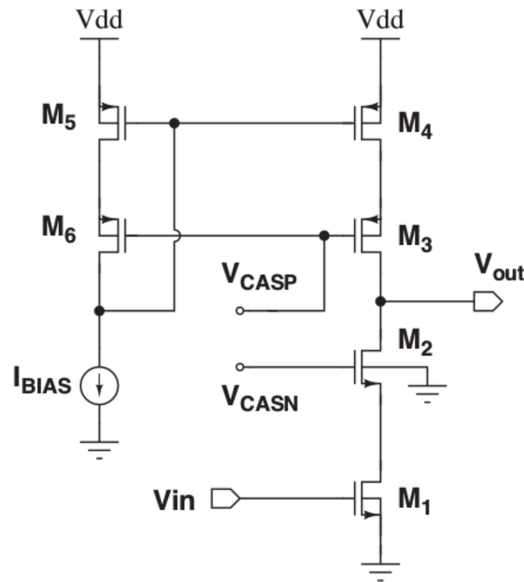


Figure 2.16: n-type telescopic cascode amplifier, taken from [8]

Telescopic cascode amplifier (fig.2.16)

With the *telescopic cascode amplifier* the goal is to further increase the gain by cascoding the load part as well. The added resistance to the path to V_{DD} can be calculated in the same way as it was done before.

The resulting gain is:

$$\begin{aligned} A_v &= -g_{m1}(r_{0casN}//r_{0casP}) \\ &= -g_{m1}[r_{01} + r_{02} + (g_{m2} + g_{mb2})r_{01}r_{02}]//[r_{03} + r_{04} + (g_{m3} + g_{mb3})r_{03}r_{04}] \end{aligned} \tag{2.49}$$

The same considerations discussed for the simple cascode configuration on its operating conditions in terms of currents and voltages hold for this topology.
By taking advantage of this topology the gain can be made easily grater than $30dB$, making it a good candidate for closed loop configuration systems.

2.4.3 Differential Amplification Stages

Differential stages are needed in order to generate signals proportional to other signals difference.
In order to simplify the following reasoning a differential signal will be decomposed in two components:

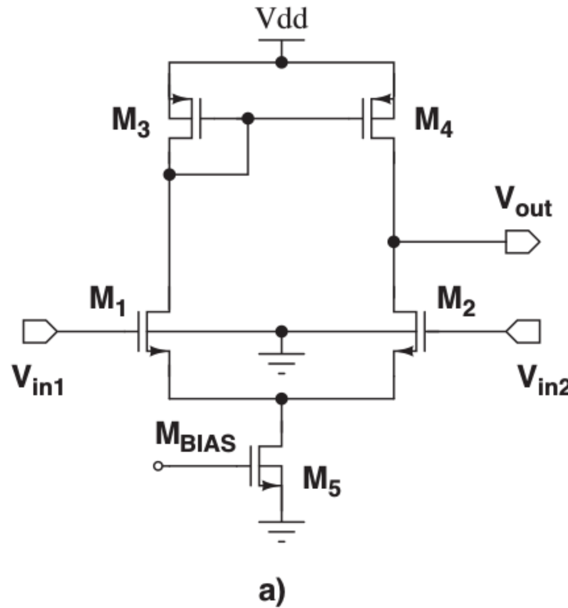


Figure 2.17: n-type differential cell, taken from [8]

– Differential mode (DM)

$$X_{DM} = X_2 - X_1 \tag{2.50}$$

– Common mode (CM)

$$X_{CM} = \frac{X_2 + X_1}{2} \tag{2.51}$$

Obviously a good differential stage needs to be very sensitive to DM while being insensitive to CM.
The purpose of the presented topologies is only to discern this components. For what concerns amplification and performances, the same techniques applied to single ended stages holds in these cases.

Differential cells with current mirror loads (fig.2.17)

The presented circuit is as current steerer: the differential signal applied to the gates of its input transistors (M_1 and M_2) distributes the current imposed by the bias transistor (M_5) between the two branches. This currents are than converted to two voltages by the mean of the loads (M_3 and M_4). When the same voltage is applied at the inputs the currents of the two branches are the same and equal to

half of the current flowing thorough M_5.

Considering the case of two perfectly symmetrical branches, we can rename the quantities of the components as: $g_{m|in} = g_{m1} = g_{m2}$ and $r_{0in} = r_{01} = r_{02}$ for the input transistors, and $r_{0L} = R_{03} = R_{04}$ for the loads.

The differential gain can be computed as:

$$\begin{aligned} A_{DM} &= \frac{V_{out2} - V_{out1}}{V_{DM|in}} \\ &= g_{m|in}(r_{in}//r_{0L}) \end{aligned} \tag{2.52}$$

The absence of contribution by M_5 is justified by the fact that a small pure differential signal will not condition the current flowing through it, thus for the small signal equivalent the two branches common node is a ground. Bearing in mind this aspect, follows naturally that the gain of this stage equals the one to which one half of the input is applied to two independent common source amplifiers.

As for the common mode, this time, the common node is no more static, therefore the transconductance is the one of a degenerated source amplifier:

$$\begin{aligned} A_{CM} &= \frac{V_{CM|out}}{V_{CM|in}} \\ &= \frac{g_{m|in}(r_{in}//r_{0L})}{1 + 2g_{m|in}r_{05}} \end{aligned} \tag{2.53}$$

As can be noted, in order to suppress the common mode r_{05} bust be made as large as possible.

The DC can be simply calculated for the series of equivalent resistors.

Therefore differential stages has the same requirements and rules for transistor sizing of the single ended ones, however there is a supplementary condition dictated by the mismatch.

In fact, removing the condition of symmetry between the two branches, part of the CM will produce DM output, this aspect is taken in account by the *common to differential mode gain*:

$$\begin{aligned} A_{CM \to DM} &= \frac{V_{out1} - V_{out2}}{V_{CM|in}} \\ &= \frac{g_{m1}r_{01} - g_{m2}r_{02}}{1 + (g_{m1} + g_{m2})r_{05}} \end{aligned} \tag{2.54}$$

So the effect of mismatch grows with the gain of the cell, and can be reduced together with common mode gain.

Chapter 3

Front-End Architecture

This chapter will focus on the studied analog-front end, describing its operating principle, the chosen design implementation, transistors sizing and optimization. Evaluation of estimated performances are also reported with a reference to the simulations results of chapter 4.

The biasing network is not reported since is formed by standard blocks discussed in subsection 2.4.1, in the case of the adoption of a different biasing approach the subject will be addressed in the text. As previously mentioned, unlike in the rest of the front-end, this biasing blocks will be implemented off pixel and shared between a number of channels. Following this fact they where designed separately from the main circuit in perspective of the final layout.

Circuits description will be handled at high level of abstraction using small signal equivalent models. For the details on the employed models please refer to sections 2.1.2 and 2.4. In the course of the chapter time domain signals will indicated with small letters (like i and v) while the quantity in frequency domain with capital ones (for example I and V).

The described architecture was firstly implemented in a 65nm node since it is

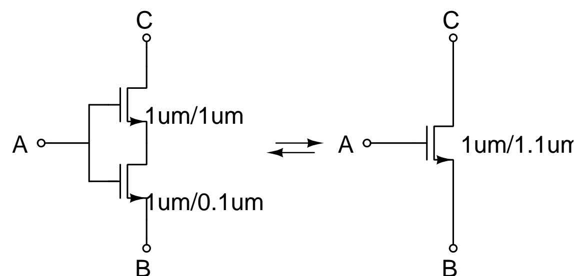


Figure 3.1: Example of an equivalent series transistor.

a consolidate technology in the filed of radiation sensors electronics compared to the novel 28nm one, so it represented both a starting point to test the feasibility of the proposed architecture and a benchmark of its performance. The 28nm implementation of design started afterward, and has currently reached the point in which the input amplifier has been completed. The porting of the 65nm design has been executed by scaling of a factor 0.5 every transistor. This was done in order to have a starting circuit that can be successively optimized. The regular fabrics imposed by the manufacturer, due the new lithography techniques, requires to have transistors of the same L in the design in order to save area. As discussed in 2.3.4, two dummy transistors must be placed for every transistor with a different L. To maintain a good scaling from 65nm to 28nm the design was divided in two major blocks on the basis of the L value: 100 nm and 1 μm. The 0.5 scale factor was maintained exploiting the fact that two transistors in series with the gate terminals shorted together are equivalent to one transistor with an L equal to the sum of the L of the two. This concept is presented in fig.3.1. When the 0.5 scaling didn't allow to satisfy the regular fabrics rule, the value of L was rounded.

In the course of this chapter characteristics of both implementation extracted from simulations will be presented when available, in any case evaluation based on small signal models will be carried over. The necessary parameters were extracted from technology test simulations presented in 4.1.

In the first part of the chapter an overview of the full chain is presented, while the discussions on the specific blocks will be presented afterwards with a top-to-bottom approach.

3.1 Overview

The block scheme of the full front-end chain is presented in fig.3.2. The system is composed by a first amplifier which is directly connected to the sensor and converted to voltage signals the input current pulses. The amplifier also features DC current compensation. The formed signal is then compared with a leading edge discriminator to a threshold level V_{thr}, a digital 1 signal is produced, as result, with its rising edge synced with the threshold crossing event. An offset correction circuit is added in order to equalize channel to channel non-uniformities and to impose a desired DC level V_{bl} to the discriminator input.

The selected amplifier architecture was chosen due to its good timing properties and low noise. The choice of simple leading edge discriminator was driven by the desire to quantify the impact of time-walk and then decide the approach to eliminate

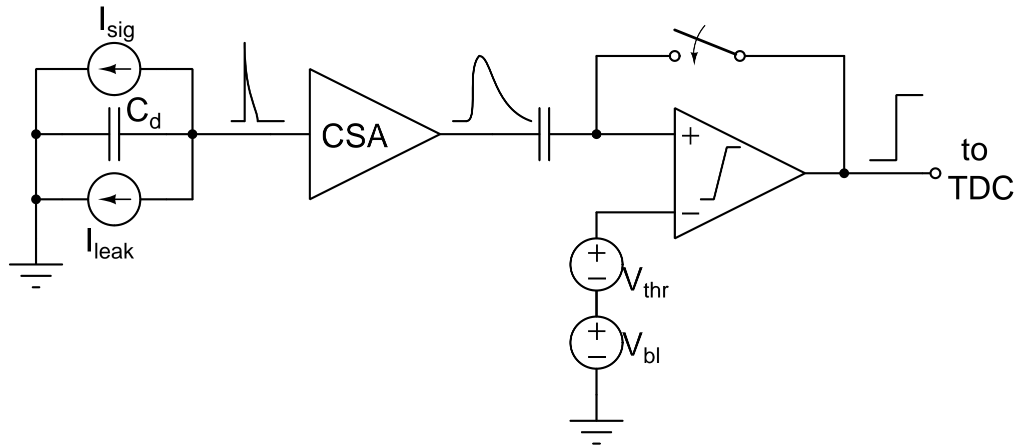


Figure 3.2: Block scheme of the implemented front-end, from left to right: sensor model, charge sensitive amplifier and discriminator.

it with a constant fraction discriminator or to correct it with a time-over-threshold measure. In any case this simple discriminator topology constitutes a basic building block for more complex ones (discriminator architectures are discussed in 1.3.2).

The sensor is modelled by two current sources in parallel to a capacitor: the signal has been split int two components in order to highlight the presence of the DC leakage current I_{leak} superimposed to the signal I_{sig}. The ideal capacitor C_d simulates the sensor own capacitance, details on sensor working principle and characteristics of the specific ones with whom the amplifier will be connected were presented in 1.2.1.

3.2 Charge Sensitive Amplifier

This section starts describing the operation of the input amplifier observed from the outside, aspects discussed later will be used in advance.

A *Charge Sensitive Amplifier (CSA)* is a circuit in which the voltage amplitude at its output is proportional to the total charge forming the input current signal. As discussed in 1.2.1 this is a very useful characteristics for radiation sensors since the charge relative to their current signals is proportional to the energy deposited by the radiation inside the material. Another important feature of this circuit is the fact that it exhibits almost constant peaking and falling time, with its related consequences for time-walk correction discussed in 1.3.1.

It consists in an amplifier with a capacitor C_f and a resistor R_f in negative feed-back, in the actual design R_f is an equivalent active element provided by the

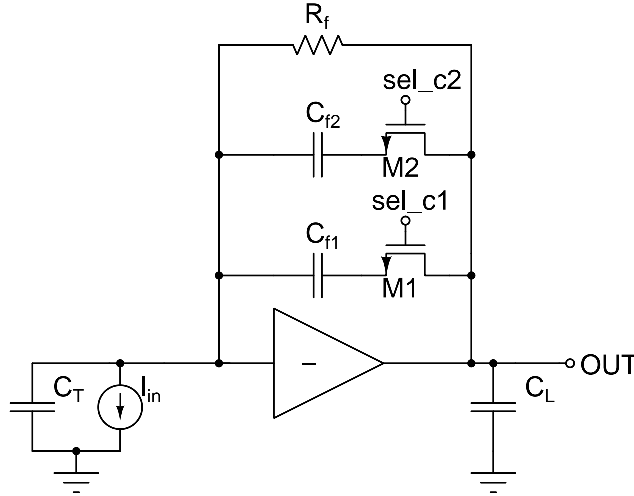


Figure 3.3: Charge sensitive amplifier.

Krummenacher filter, while C_f is a combination of two parallel capacitors selected with two switches, implemented with transistors as showed in fig.3.3. The size of this two transistors and the capacitance values of the capacitors are reported in tab.3.1.

Its operation will be explained considering the ideal case in which:

- the core amplifier has infinite gain $A_0 \to \infty$
- the *CSA* ha no bandwidth limitation
- the feedback path is only capacitive: with a resistance $R_f \to \infty$
- the input signal is delta shaped $i(t) = \delta(t)$

The input node can be considered a virtual ground, so the instantaneous current signal is totally integrated by C_f. In this condition the *CSA* behaves as an ideal integrator:

$$\begin{aligned} v(t)_{out} &= \frac{1}{C_f} \int_0^\infty i_{in} \delta(t) dt \\ &= \frac{Q_{in}}{C_f} u(t) \end{aligned} \tag{3.1}$$

It's response is a step $u(t)$ voltage with an height proportional to the input charge by $\frac{1}{C_f}$. Considering the maximum capacitance of 7.5 fF a gain charge-to-voltage of $G_c = 133 mV fC^{-1}$ is obtained.

CSA Transistors Sizing		
	$\left(\frac{W}{L}\right)\left[\frac{\mu m}{\mu m}\right]$	
	65 nm	28 nm
M_1	$\frac{1}{0.08}$	$\frac{1}{0.1}$
M_2	$\frac{1}{0.08}$	$\frac{1}{0.1}$
CSA Capacitors		
	C $[fF]$	
C_{f1}	2.5	
C_{f2}	5	

Table 3.1

Considering now the case of a core amplifier with finite gain A_0, the input node can no more be considered a virtual ground, in this case the new nodal equation is:

$$I_{in}(s) + V_{in}(s)sC_T + (V_{in}(s) - V_{out}(s))sC_f = 0 \tag{3.2}$$

Where C_T is the total capacitance to ground seen by the input node, in this case the contribution are from the sensor C_s and the gate-source capacitance of the input transistor $C_{gs|in}$:

$$C_T = C_s + C_{gs|in} \tag{3.3}$$

Using the definition of open loop gain $A_0 = -\frac{V_{out}}{V_{in}}$ (the minus sign is used to explicit the inverting behaviour of the core amplifier) the CSA output becomes:

$$V_{out}(s) = \left(\frac{A_0}{C_T + (1 + A_0)C_f}\right)\left(\frac{I_{in}(s)}{s}\right) \tag{3.4}$$

Returning in the time domain:

$$v_{out}(t) = \left(\frac{A_0}{C_T + (1 + A_0)C_f}\right)Q_{in}u(t) \tag{3.5}$$

As can be noted the effect of a finite gain is loss in the total voltage to charge gain. The dependence from C_T comes from the fact that the input node is no more a virtual ground: part of the input current will now charge this capacitance instead

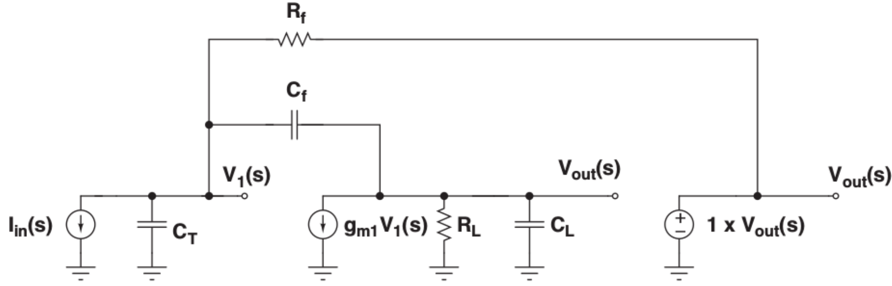


Figure 3.4: CSA simplified equivalent circuit.

of the feed-back one, reducing the charge to voltage gain. Thus the ideal case is approximated for both the condition $A_0 \gg 1$ and $(1+A_0)C_f \gg C_T$. Using the values obtained with the core amplifier simulations(A_{065} =27.6 dB and A_{028} =29.9dB), and taking $C_T = 100fF + 5fF$, this amplitude loss can be quantified as 2.4% for 65nm and 1.5% for 28 nm. The effect of finite bandwidth will now be considered. The transconductance transfer function can be evaluated considering the equivalent circuit in fig.3.4, its derivation can be found in [8], here a simplified version is used:

$$T(s) = \frac{R_f}{s^2 \frac{\zeta R_f}{g_{m|in}} + sR_fC_f + 1} \tag{3.6}$$

Where:

$$\zeta = C_TC_L + C_f(C_T + C_L) \tag{3.7}$$

In the hypothesis in which the feedback resistance is infinite, the relation can be simplified in :

$$R_f \gg 1 \Rightarrow T(s) \simeq \left(\frac{1}{sC_f}\right)\left(\frac{1}{1+s\tau_r}\right) \tag{3.8}$$

Considering again a delta-shaped signal and using the inverse Laplace transform we can observe this effect in the time domain:

$$v_{out} = \frac{Q_{in}}{C_f}\left(1 - e^{-\frac{t}{\tau_r}}\right) \tag{3.9}$$

Where the substitution was:

$$\tau_r = \frac{\zeta}{g_{m|in}C_f} \tag{3.10}$$

The first effect of BW limitation is a slow down of the output signal, which now exhibits a finite rising time. τ_r depends mainly on capacitive elements that the designer can't control, like the sensor and load capacitance. Anyway $g_{m|in}$ can be

increased in order to control this term, but generally this will increase the input transistor C_{gs} limiting , at a certain point, the control on τ_r due to its contribution to C_T. Using the value of $g_{m|in}$ obtained from the fit presented in tab.4.1 ($48.6\mu S$ for 65n and $55.4\mu S$ for 28nm) τ_r results of 4.17 ns and 4.88 ns.

Adding now the fact that the feedback loop resistance is not infinite, both the two poles of (3.6) will affect the response of the circuit. In order to streamline the process, the denominator will be approximated using:

$$\begin{aligned} D(s) &= (1+s\tau_f)(1+s\tau_r) \\ &= 1 + s(\tau_f+\tau_r) + s^2\tau_f\tau_r \\ &\overset{\tau_f \gg \tau_r}{\simeq} 1 + s\tau_f + s^2\tau_f\tau_r \end{aligned} \tag{3.11}$$

Comparing this relation with (3.6), the relation for τ_f can be obtained:

$$\tau_f = R_f C_f \tag{3.12}$$

Inserting an estimated value of the feed-back resistance provided by the Krummenacher filter (discussed in the next section), τ_f results in 23ns for the 65nm design and 21ns for the 28 nm one. Has can be seen the condition $\tau_f \gg \tau_r$ results reasonable for the considered *CSA*. The new transfer function reads:

$$T(s) = \frac{R_f}{(1+s\tau_f)(1+s\tau_r)} \tag{3.13}$$

Considering for the moment only the contribution of this characteristic time, its effect is considered at both high and low frequencies:

$$s\tau_r \ll 1 \Rightarrow T(s) = \frac{R_f}{1+sC_fR_f} \quad \Longrightarrow \quad \begin{aligned} s \ll \frac{1}{R_fC_f} &\Rightarrow T(s) = R_f \\ s \gg \frac{1}{R_fC_f} &\Rightarrow T(s) = \frac{1}{sC_f} \end{aligned} \tag{3.14}$$

At low frequency most of the current flows thorough the feedback resistance, this defines the DC level of the output on the basis of the sensor leakage current. Its effect at higher frequencies is really important for a radiation sensor amplifier, considering again the Laplace anti-transform and a delta current input:

$$v_{out}(t) = \frac{Q_{in}}{C_f} e^{-\frac{t}{\tau_f}} u(t) \tag{3.15}$$

As can be seen R_f discharges C_f re-establishing the DC level after a certain time. In this way the amplifier is ready to process the next signal: R_f value must be chosen according to the expected average signal rate, in order to prevent pile-ups.

Now the combined effect of the two poles is evaluated taking the inverse Laplace transform of (3.13):

$$\begin{aligned} v_{out} &= Q_{in}\frac{R_f}{\tau_r - \tau_f}\left(e^{-\frac{t}{\tau_r}} - e^{-\frac{t}{\tau_f}}\right)u(t) \\ &= \frac{Q_{in}}{C_f}\frac{\tau_f}{\tau_r - \tau_f}\left(e^{-\frac{t}{\tau_r}} - e^{-\frac{t}{\tau_f}}\right)u(t) \end{aligned} \tag{3.16}$$

The output signal now resembles the one expected from a CSA, with a certain peaking and falling times independent on the signal. This is an important aspect for timing applications, as discussed in 1.3. The peaking time can be evaluated taking the first derivative of the pulse function and equating it to zero. The resulting time is:

$$T_{pk} = \frac{\tau_f \tau_r}{\tau_f - \tau_r}\ln\left(\frac{\tau_f}{\tau_r}\right) \tag{3.17}$$

So, for fast rising time constants, the peaking time is mostly independent on the slow exponential of τ_r, even though the whole signal is delayed by it. Inserting the values calculated above the peaking time of the considered designs is estimated to be 8.7ns for the 65nm design and 9.3ns fro 28nm one. Using this value the peak amplitude can be evaluated:

$$V_{pk} = \frac{Q_{in}}{C_f}\left(\frac{\tau_f}{\tau_r}\right)^{\frac{\tau_r}{\tau_r - \tau_f}} \tag{3.18}$$

So another effect of the two exponentials is a further reduction on the charge-voltage gain, which is called ballistic-deficit. Inserting the numbers this effect will bring the charge-gain to 89 $mVfC^{-1}$ and 84 $mVfC^{-1}$.

Finally the effect of signal shape is considered. For sake of simplicity the signal is modelled with a boxed negative exponential. This can be considered a good approximation since good sensor signals exhibit a very fast rising time compared to the falling one. The Laplace transform of the signal is considered:

$$i_{in} = I_0 e^{-\frac{t}{\tau_s}}u(t) \Rightarrow I_{in} = \frac{I_0 \tau_s}{1 + s\tau_s} \tag{3.19}$$

To make the effect more evident the signal is applied to the transfer function that

does not take τ_r into account:

$$\begin{aligned} V_{out} &= I_0\tau_s \frac{R_f}{(1+s\tau_f)(1+s\tau_s)} \\ &= Q_{in}\frac{R_f}{(1+s\tau_f)(1+s\tau_s)} \end{aligned} \tag{3.20}$$

So the falling time of the signal adds another pole to the expression for the output, determining, similarly to τ_r, a ballistic deficit. This last effect has repercussion in timing: due to the statistical nature on the signal generation inside the sensor, variation in the signal shape are inevitably seen at the *CSA* output as source of uncertainty.

3.2.1 Core Amplifier

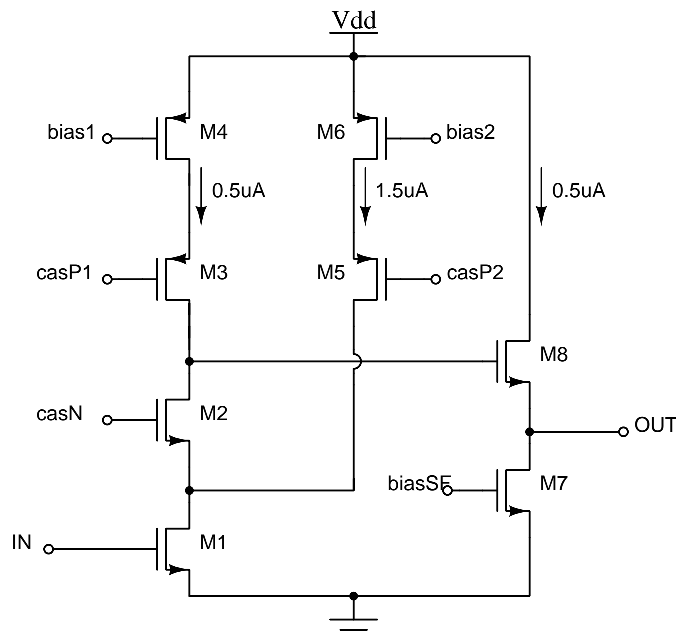


Figure 3.5: Telescopic cascode amplifier with split bias current branches.

The core amplifier of the *CSA* is a telescopic cascode amplifier with split bias current branches, it has a n-type input transistor followed by an n-type source follower used as buffer. The schematic of the circuit is presented in fig.3.5 while its transistors sizing reported in tab.3.2.

This amplification stage operates similarly to the telescopic cascode configuration described in 2.4.2, with the main difference that it is optimized for low power

Core Amp Transistors Sizing		
	$\left(\frac{W}{L}\right)\left[\frac{\mu m}{\mu m}\right]$	
	65 nm	28 nm
M_1	$\frac{8}{0.2}$	$\frac{4}{0.1}$
M_2	$\frac{2}{0.1}$	$\frac{1}{0.1}$
M_3	$\frac{2.5}{0.25}$	$\frac{1.5}{0.1}$
M_4	$\frac{0.5}{2}$	$\frac{0.25}{1}$
M_5	$\frac{2.5}{0.25}$	$\frac{1.5}{0.1}$
M_6	$\frac{0.5}{3}$	$\frac{0.25}{1.5}$
M_7	$\frac{0.5}{6}$	$\frac{0.3}{3}$
M_8	$\frac{4}{0.2}$	$\frac{2}{0.1}$

Table 3.2

consumption: the two bias branches have respectively $0.5\mu A$ and $1.5\mu A$ of bias current. Adding to these currents only $0.5\mu A$ for the source follower results in an estimated power consumption of $2.5\mu W$ in the 28nm case.

The biasing of the first stage is provided with a variation of the usual current mirror: in this biasing circuit the references voltages are provided, likewise in the normal current mirror, at the gate of a transistor. The main difference is in the fact that the device which is connected to the current source is no more a single transistor but an entire replica of the circuit to be biased.

The bias branches forms two p-type cascodes: each of them has its equivalent resistance given by (2.46), in this way we get for the two path:

$$\begin{aligned} r_{casP1} &= r_{03} + r_{04} + g_{m3} r_{03} r_{04} \\ r_{casP2} &= r_{05} + r_{06} + g_{m5} r_{05} r_{06} \end{aligned} \tag{3.21}$$

The n-cascode resistance can be evaluated in the same way, but this time the total resistance towards the small signal ground is the parallel between the one of M_1 and the one of the second bias branch:

$$r_{casN} = (r_{01} // r_{casP2}) + r_{02} + g_{m2}(r_{01} // r_{casP2}) r_{02} \tag{3.22}$$

This two resistance, similarly to the case of the normal telescopic cascode amplifier, has a total load resistance which is the parallel of this two contributions. The transconductance of this stage can be calculated as in the case of simple cascode configuration (eq.(2.47), using as the resistance the parallel of M_1 and the second p-cascode instead of M_1 alone:

$$\begin{aligned} G_m &= g_{m1} \frac{(r_{01}//r_{casP2}) + g_{m2}(r_{01}//r_{casP2})r_{02}}{(r_{01}//r_{casP2}) + r_{02} + g_{m2}(r_{01}//r_{casP2})r_{02}} \\ &\simeq g_{m1} \end{aligned} \tag{3.23}$$

The total gain of stage is simply:

$$A_0 = -G_m(r_{casN}//r_{casP1}) \tag{3.24}$$

The value of this gain extracted from simulations results in 27.6 dB for the 65nm design and 29.9 dB for the 28 nm one.

The input transistor sizing was carried out using the rules described in 2.4.2: the maximum value of $\left(\frac{W}{L}\right)$ was decided on the basis of the saturation of $g_m(W/L)$, in fact at a certain point g_m cannot be increased significantly even doubling $\left(\frac{W}{L}\right)$, which will lead to bigger gate capacitances. In fact with $\left(\frac{W}{L}\right)$ the g_m value obtained using eq.(2.23) with the fit parameters obtained in sec.4.1 is 55.3 μS for the 28nm design, doubling for instance W will result in 57.5 μS, while the capacitance will double.

In the frequency domain the circuit can be approximated, due to the high impedance of the cascode output node, to one with a single pole:

$$A_0(s) = -\frac{A_0}{1 + \frac{s}{\omega_p}} \tag{3.25}$$

where the pole angular frequency is give by:

$$\omega_p = \frac{1}{2\pi(r_{casN}//r_{casP1})C_0} \tag{3.26}$$

Where C_0 is the capacitance between the input and output node. From simulations results the bandwidth results of 13 MHz for the 28 nm design.

The last consideration is on the driving strength of the source follower, its output

resistance can be computed as:

$$r_{out} \simeq \frac{1}{g_{m8}} \tag{3.27}$$

Using the fit parameters its value results in $\sim 80\ k\Omega$.

3.2.2 Krummenacher Filter

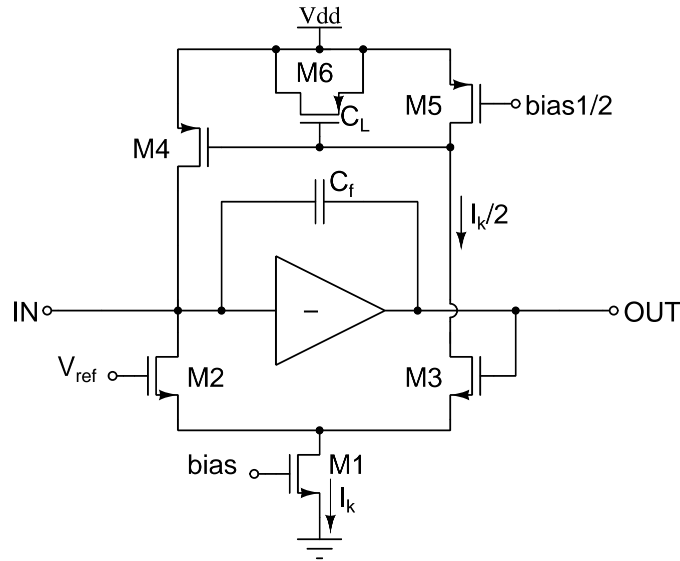


Figure 3.6: Krummenacher filter transistors level schematic. The gate of M_3 is actually connected to the output of the amplification stage, before the buffer, this detail was omitted in this schematic for simplicity reasons. Transistor M_6 is indicated as C_L since it is used as a capacitor.

The Krummenacher filter is represented in fig.3.6 with its transistors sizing presented in tab.3.3. This circuit has been introduced for three purposes:

– Compensate the DC leakage current sunk from input by the sensor.

– Provide DC voltage levels to both the input and the output of *CSA*, configurable via a reference voltage V_{ref}.

– Provide a current which discharges the feedback capacitor C_f, constituting an effective feedback resistance R_f

The circuit consists in a modified differential pair in which the total current I_k is fixed by the biasing transistor M_1, in the right branch the current is imposed

Krummenacher Filter Transistors Sizing		
	$\left(\frac{W}{L}\right)$ $\left[\frac{\mu m}{\mu m}\right]$	
	65 nm	28 nm
M_1	$\frac{0.1}{4}$	$\frac{0.1}{4}$
M_2	$\frac{1.2}{1}$	$\frac{0.6}{0.5}$
M_3	$\frac{1.2}{1}$	$\frac{0.6}{0.5}$
M_4	$\frac{0.5}{3}$	$\frac{0.3}{1.5}$
M_5	$\frac{0.5}{3}$	$\frac{0.3}{1.5}$
M_6	$\frac{8}{7}$	$\frac{4}{4}$

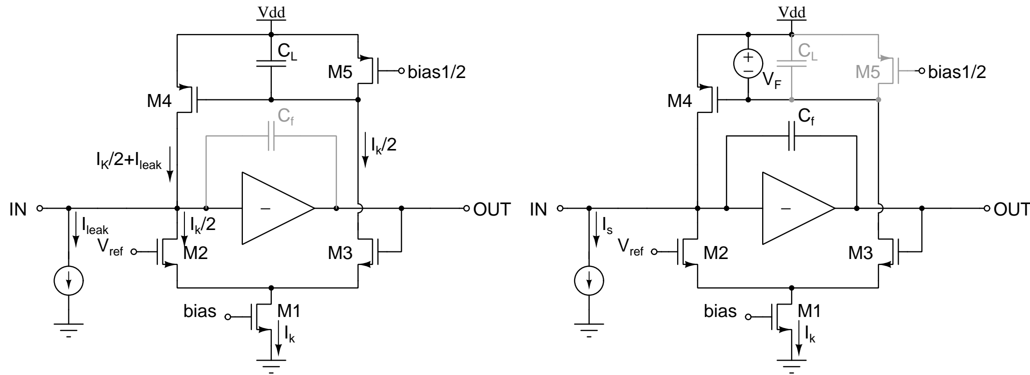
Table 3.3

to be half of I_k by M_5. The right branch of the pair is connected to the input of the amplifier, the gate of the transistor M_3 is connected to the output of the core amplifier amplification stage, before the buffer. The value of this current must be chosen based on the magnitude of the leakage current that needs to be compensated, since $\frac{I_k}{2}$ is the maximum current that can be provided by the input branch. This current is configurable within the limits posed by the stability of the circuit, but in the case of 3D sensors, only few nA need to be compensated since this is the expected leakage current.

The role of M_2 is to define the DC voltage point at the input via V_{ref}, since at equilibrium its current is locked to $\frac{I_k}{2}$.

The operation of the circuit changes in relation to the frequency of the input current signal. In order to understand the properties of the circuit, it is decomposed in two independent paths, illustrated in fig.3.7, relative to two opposite cases: a DC current and an high frequency one:

– Low frequency behavior (fig.3.7a): In this condition no current can flow through C_f. If no DC current is present the differential pair is balanced, thus the current is $\frac{I_k}{2}$ in both branches. But if sensor is leaking a current I_{leak}, this amount of current is sunken from the left branch, lowering the input node voltage. This imbalance is sensed by the amplifier which drives the output node, and therefore the gate of M_3, up. In turn M_3 rises the right branch current which discharges C_L, decreasing M_4 gate voltage, which now can provide more current to the input branch. At the equilibrium the current provided



(a) Low frequency path: C_f can be considered as an open circuit from low frequency signals.

(b) High frequency path: C_L is considered a short circuit for high frequency variations, locking the voltage difference across it as the stored one V_F during DC operation.

Figure 3.7: Krummenacher filter representation of two independent paths based on frequency of operation. M_6 is represented as a capacitor C_L for sake of clarity.

by M_4 equals $\frac{I_k}{2} + I_{leak}$, compensating it. A consequence of this fact is that the input and output voltage DC levels will no more coincide. In order to better understand this mechanism the small signal transfer function of this current-to-voltage path is presented:

$$\begin{aligned} Z_{lf} &= \frac{V_{out}}{I_{in}} \\ &= \frac{V_{out}}{I(M_4)} \\ &= \frac{V_{out}}{g_{m4}\left(\frac{g_{m3}}{2}V_{out}\right)\left(\frac{1}{sC_L}\right)} \\ &= \frac{2sC_L}{g_{m3}g_{m4}} \end{aligned} \tag{3.28}$$

The effect of the circuit at low frequencies is described by the zero placed at $s = 0$, which creates a high-pass filter that reduces to zero the total gain at DC level. C_L has a dual role: it drives M_4 gate and decide the frequency at which the DC current signal starts to be attenuated ($s = g_{m3}g_{m4}(2C_L)^{-1}$).

– High frequency behavior (fig.3.7b): in this case the current signal I_s charges C_f as in the ideal *CSA* described at the beginning of the section. When the

output voltage rise , the current in the right branch is increased through M_3. This time any fast AC variation is absorbed by C_L and compensated by V_{DD}, in this way the current in the input branch is no more set by M_4 since its gate voltage is fixed to the value V_F loaded in C_L during the low frequency operation. Therefore the current variation thorough M_3 must be mirrored by M_2 in the symmetric branch of the differential pair:

$$\begin{aligned} Z_{hf} &= \frac{V_{out}}{I_{in}} \\ &= \frac{2}{g_{m3}} \end{aligned} \tag{3.29}$$

It must be noted that only half of g_{m3} contributes to this feed-back impedance since this is the signal applied to a differential pair driven in single-ended mode. So it results that Z_f ha a resistive behaviour and act as the feed-back resistance of the *CSA*: so it can be assumed for the small signal at high frequency that $Z_f \equiv R_f$. Considering a current of $25nA$ for each branch and the sizing of M_3, the value of this resistance is $\sim 3M\Omega$.

3.3 Leading Edge Discriminator

The block-scheme level representation of the complete leading edge discriminator is presented in fig.3.8, its transistors sizing is in tab.3.4. The system is formed by two amplification stages (described in the next section) followed by an inverter which serves both to generate a positive signal after the inverting amplification and as a digital buffer. An offset correction circuit is also added in order to both set the desired DC level of the input and to compensate the presence of voltage offset in threshold setting, details on the mechanism of this circuit are discussed in 3.3.2. In fig.3.9 the transistor level implementation is presented.

3.3.1 Amplification Stages

A discriminator is a device which produce a digital signal once the input signal as crossed a certain value V_{thr}. A good timing discriminator is the one in which the delay time t_d between the threshold crossing and the rising edge of the digital signal is insensible to the slope of the original signal at the threshold crossing, or to other timing independent variabilities. This obtained using an amplifier which drives an inverter which generates the digital pulse once its input voltage reaches a certain

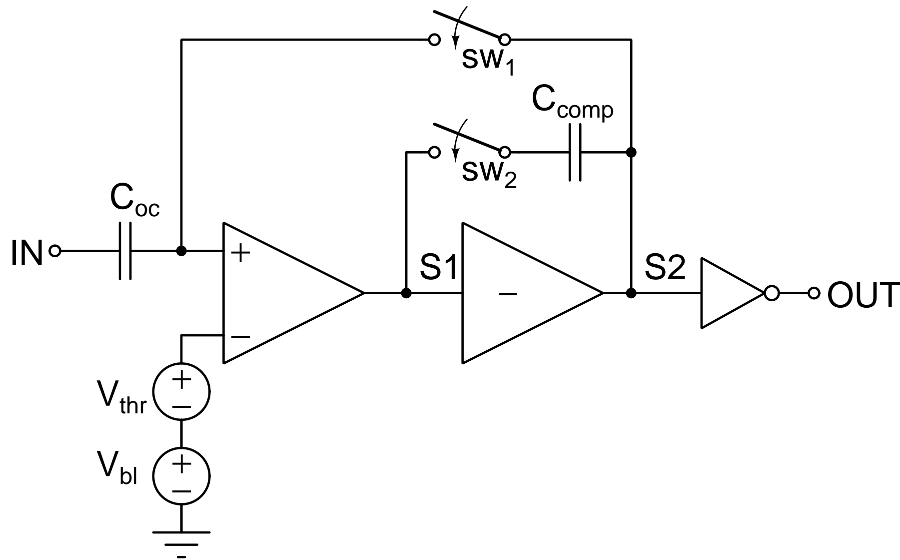


Figure 3.8: Block scheme of the discriminator complete with offset correction circuit. From left to right: first differential amplification stage, second single ended stage, inverter. Note that the inverting terminal of the first stage corresponds to the non-inverting terminal of the full amplification chain.

value V_{inv}. In order to obtain good results different conditions must be met:

1. The amplification stage must have a gain as high a possible. Supposing an infinite gain stage, two signals with different slopes amplified by it will produce two output vertical edges which will cross V_{inv} at the same time. Moreover assuming that V_{inv} varies from an inverter to another, a infinite steep slope will cancel this effect.

2. The amplification stage needs an high bandwidth as possible, otherwise the output signal edge will be smoothed becoming a bad timing reference for the following inverters for the same reason discussed above.

3. The DC level of the output when the input signal is at threshold level must coincide with V_{inv}. In absence of this condition an offset will be present to the desired V_{thr}. In a real case mismatch variations of both the inverter and the amplifier will make achieving this condition impossible.

Achieving both the 1 and 2 ideal conditions simultaneously or the third one, makes the discriminator delay time t_d zero. Since none of this conditions is achievable in a real case, every discriminator will inherit an intrinsic dependence on its

mismatch variations and on signal slope at threshold crossing. This conditions can be mitigated by optimizing the discriminator design in all of this aspects.

Another source of uncertainties is the noise introduced by the amplifier itself, but its a secondary effect since the amplified signal presents an high slope.

A two stage configuration was selected due to the difficulty to implement a single high gain differential stage: to obtain an high gain a larger current needs to be put in both the two branches, or a more complex topology (like a cascode one) has to be implemented twice. For this reason a low gain differential input stage was chosen in order to sens the differential threshold crossing, while a second high gain single ended stage was used to produce a sharp signal. The first stage is a p-type inputs since they operates well at voltages closer to zero, the base line of the input signal must be set in this region in order to obtain a large input range.

The gain of the first stage is the one of the differential cell:

$$A_{01} = g_{m1}(r_{02}//r_{04}) \tag{3.30}$$

The gain of this stage could be increased acting on $\left(\frac{W}{L}\right)$ of the input transistors, leading however to a grater t_d because of the gate capacitance. Attention was paid to the DC level of the output node S_1 when both input terminals are at V_{thr}: this level must lie inside the high gain input range of the DC transfer function of the next stage, otherwise the full second stage gain will not be fully exploited. In this condition, a variation on this node voltage will create a signal dependent dead time between the activations of the two stages (the second stage starts to amplify at full gain under 800mV). A better solution would have been to put the DC voltage of V_{S1} at the input level which correspond to an output value equal to V_{inv}, but this was not possible with realistic transistors sizing: the chosen approach was to share the fulfillment of this condition to the sizing of the second stage and the inverter. The total bias current of this stage is $2\mu A$ ($1uA$ per branch) with a $g_{m1} = 25.5\mu S$, leading to a gain of ~ 10.

For the second stage a cascode configuration was used due to its higher gain, calculated as:

$$A_{02} = -g_{m6}(r_{casN}//r_{08}) \tag{3.31}$$

Where r_{casN} is calculated as the usual cascode resistance:

$$r_{casN} = r_{06} + r_{07} + g_{m7}r_{06}r_{07} \tag{3.32}$$

This led to a gain of ~ -90 , with $1\mu A$ of bias current with the same g_m of the first

stage. The total gain of the amplifier can be expressed by the product of the ones of the respective stages.

$$A_{disc} = \frac{A_{01}A_{02}}{(1+s\tau_1)(1+s\tau_2)} \tag{3.33}$$

Resulting in a total gain of 29.5 dB.

Finally the inverter sizing was chosen in order to have a $V_{inv} = 400mV$, and this results in the big gate area that can be found in tab.3.4. The effect of the relative big gate capacitances introduced by them will not be a concern for timing since they will simply increase the propagation delay of this unit (the loading effect on the second amplification stages are negligible since their value would be of few fF).

The final t_d obtained by the simulation is 5 ns.

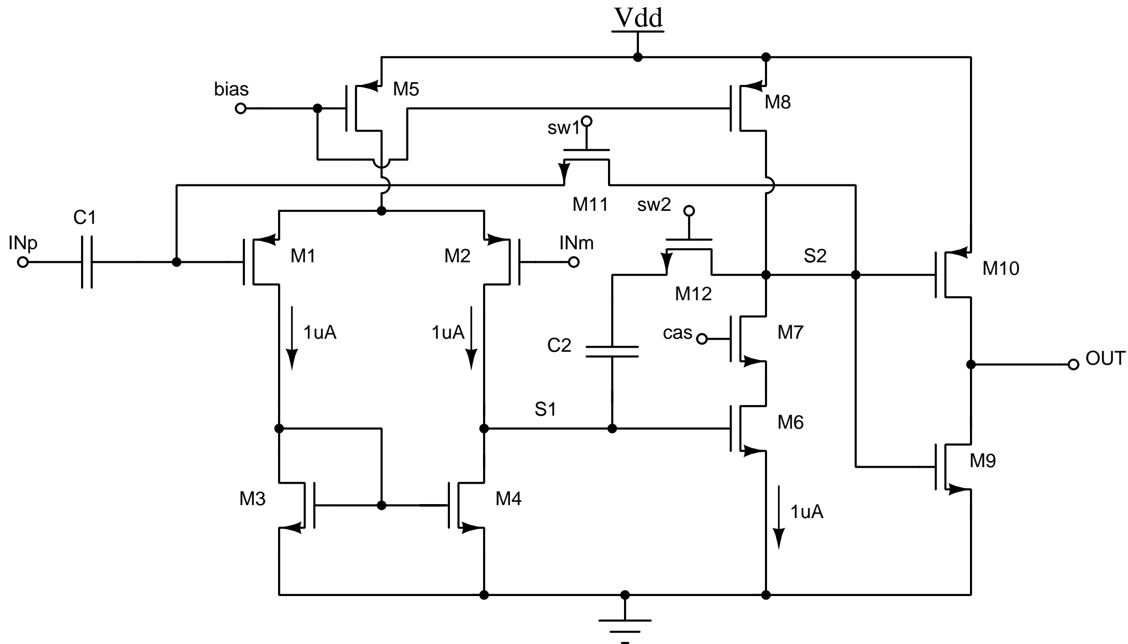


Figure 3.9: Transistors level implementation of the leading edge discriminator with offset correction.

3.3.2 Offset Correction

As described in the previous section, the effective threshold V_{thrE} will differ from the one imposed on the non-inverting terminal of the amplifier V_{thr} by a certain unknown offset value V_{os}.

$$V_{thrE} = V_{thr} + V_{os} \tag{3.34}$$

Discriminator Transistors Sizing	
	$\left(\frac{W}{L}\right)\ \left[\frac{\mu m}{\mu m}\right]$
	65 nm
M_1	$\frac{4}{0.1}$
M_2	$\frac{4}{0.1}$
M_3	$\frac{0.7}{0.7}$
M_4	$\frac{0.5}{0.7}$
M_5	$\frac{4}{0.2}$
M_6	$\frac{4}{0.1}$
M_7	$\frac{0.35}{0.1}$
M_8	$\frac{2}{0.2}$
M_9	$\frac{1.5}{0.1}$
M_{10}	$\frac{6}{0.1}$
M_{11}	$\frac{1}{0.1}$
M_{12}	$\frac{0.2}{0.2}$
CSA Capacitors	
	C $[fF]$
C_{f1}	106
C_{f2}	35

Table 3.4

Moreover this value will vary due to mismatch and process variations. A careful design of the DC levels aware of this variation is impractical and would results in large and slow transistors.

For this reason the offset compensation discrete time circuit presented in fig.3.8 was implemented: it is composed by two capacitors C_{oc} and C_{comp} and two switches SW_1 and SW_2. In fig.3.9 the transistors level implementation is presented: as can be seen SW_1 and SW_2 where realized using two $NMOS$ transistors M_{11} and M_{12} respectively. Their bulk contacts, not represented in the schematic, were connected to ground in order to obtain a smaller current when the transistor is switched off (exploiting the V_{th} increase due to bulk effect).

This circuit can even offset the DC input level (base line) at a desired value V_{bl}.

The offset correction is performed by first measuring the offset and then storing its value inside C_{oc}, this value is thus added to the signal in normal operation. This operation will be described in steps.

1. In the inverting terminal there's no signal, so the voltage is at its baseline level V_{in0}, both switches are open and in the non-inverting terminal the desired baseline voltage V_{bl} is presented.

2. SW_2 is closed in order to prevent instabilities of the circuit (discussed below)

3. SW_1 is closed creating an unit gain negative feedback loop across the amplifier. In this condition the amplifier will act as a voltage buffer probing the difference between its input terminals. The difference will be:

$$\Delta V_{in} = V_{in+} - V_{in-} = V_{in0} - V_{bl} + V_{os} \tag{3.35}$$

The loop will now feedback this quantity with the relation:

$$\Delta V_{oc} = \frac{A_{disc}}{1 + A_{disc}} \Delta V_{in} \tag{3.36}$$

This signal will be stored inside C_{oc} and added to the input baseline V_0. The voltage difference between this two levels is now:

$$\begin{aligned} V_{err} &= V_{in0} + V_{oc} - V_{bl} + V_{os} \\ &= \Delta V_{in} - \frac{A_{disc}}{1 + A_{disc}} \Delta V_{in} \\ &= \left(\frac{1}{1 + A_{disc}} \right) \Delta V_{in} \end{aligned} \tag{3.37}$$

Which quantifies the correction error due to the gain limitation of the amplification stage, in fact considering $A_{disc} \to \infty$ this voltage difference will go to zero.

4. SW_1 is opened bring back the circuit in discrimination mode. If the switch transition is enough fast the operation will have the minimum consequences on the voltage difference loaded across C_{oc}.

5. At the non-inverting input the signal is raised to the desired threshold value above the base line ($V_{bl} + V_{thr}$). Without this operation the circuit would have been on the edge of the discrimination level and therefore it would have been triggering on noise signals.

6. SW_2 is opened since the circuit can now be stable without the presence of C_{comp}, which would have only slow down the discriminator operations. In this state the discriminator could accept the fast signals coming from the *CSA*, which will pass unchanged if C_{oc} is enough small.

Assuming to put $V_{bl} = V_{in0}$, and to have reasonable offset value $V_{os} = 10mV$ an error of $V_{err} = 10\mu V$ is expected with the gain mentioned in the previous section. Considering a slope of $\frac{V_{pk}}{T_{pk}} = 11mVns^{-1}$, this corresponds to a t_d variation $\Delta t_d = 1ps$ for the nominal case.

The stability issues mentioned in the steps 2 and 6 come from the fact that a system with two poles so closed together would exhibits an excessive ringing when closed in a unitary gain loop. C_{oc} was introduced in the system as frequency compensation technique, shifting down the frequency of the dominant pole introduced by M_6 and therefore separating the two poles. In typical applications a minimum value of this capacitance is researched due to its effect on the circuit frequency performance, but in this case, since it is disconnected during the discriminator normal operation, a minimum value of it is researched in order to occupy the smaller possible area.

The offset correction operation must be repeated because of eventual discharges of C_{oc} or long time fluctuations on V_{0in} that demand re compensation of the base line. This operations will be addressed in the future using dedicated logic to manage the delicate switching operation.

The area occupation of these integrated capacitors could be large since their capacitance per unit area is $C_A \simeq 2.12fF\mu m^{-1}$, for this reason the values of C_{oc} and C_{comp} are kept as small as possible (corresponding to $50\mu m^2$ and $16\mu m^2$, compared

to the area of $8\mu m^2$ of the largest transistor gate in the 65nm design and $2\mu m^2$ in 28nm).

Chapter 4

Simulation Results

In this chapter simulation results on the present state of front-end design proposed in chapter 3 are presented. Both tests of the 65 nm and 28 nm technology nodes, when present, are reported.

The first part of the chapter presents (in section 4.1) the preliminary tests conducted on the two technologies in order to evaluate the differences on the electrical parameters. This was done in order to have a general direction in the porting of the architecture from a technology to another, in this way the feasibility of the porting of a particular topology can be evaluated in advance.

In sections 4.2 and 4.3 characteristics and performance of the *CSA* and *Leading Edge Discriminator* are presented, analysis on the correct operation of their main components are also reported.

Finally section 4.4 illustrates the preliminary tests on the timing performances of the studied front-end, consisting mainly in tests of the 65 nm version of the design. These tests takes into account the non ideal response of the circuits with their uncertainties such as: the noise induced jitter, the channel to channel mismatch, process variation and the impact of the input signal variations.

4.1 Comparative Technology Tests

This section presents technologies tests conducted on the single transistor in order to evaluate important parameters used in the preliminary study of analog *CMOS* circuits design. As mentioned in section 2.3, the operation of scaling transistors to a smaller feature size will inevitably change their electrical parameters.

Here in particular the change of g_m in both *NMOS* and *PMOS* transistors is illustrated, since it plays an important role in the design of amplifiers and active feed-back circuits. Another two parameters which have been explored are V_{th} of the transistors and their V_{DS} when in saturation. The first parameters is useful to evaluate the DC levels of the circuits, while the second, compared with the power-supply, gives an idea on the ease of keep a vertical transistors structure in saturation.

All the simulation where performed on single diode connected transistors (illustrated in fig.4.1): this kind of connections allows to verify the dynamic properties of transistors while in saturation, acting on the imposed I_{DS} current, this condition can be matched with the ones of the transistors inside the circuits (in terms of level of inversion).

All tests are conducted with $I_{ds} = 1\mu A$ since it is an average value of the ones found in the studied design. The sizing of the transistor is: $L = 100nm$, since it represents the minimum length present in the 65 nm design; and variable W.

The same value were used for both *PMOS* and *NMOS* transistors of both technologies, the value of W was swept between 200nm and $8\mu m$ for the 65nm models, and from 200n to $6\mu m$ for the 28nm ones.

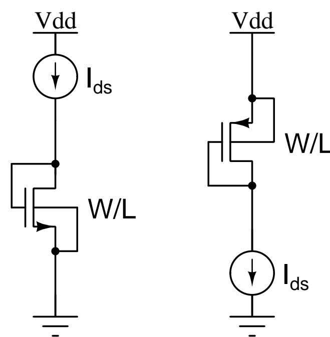


Figure 4.1: Diode connected *NMOS* and *PMOS* transistors. This configuration force them in saturation region.

Gate Transconductance

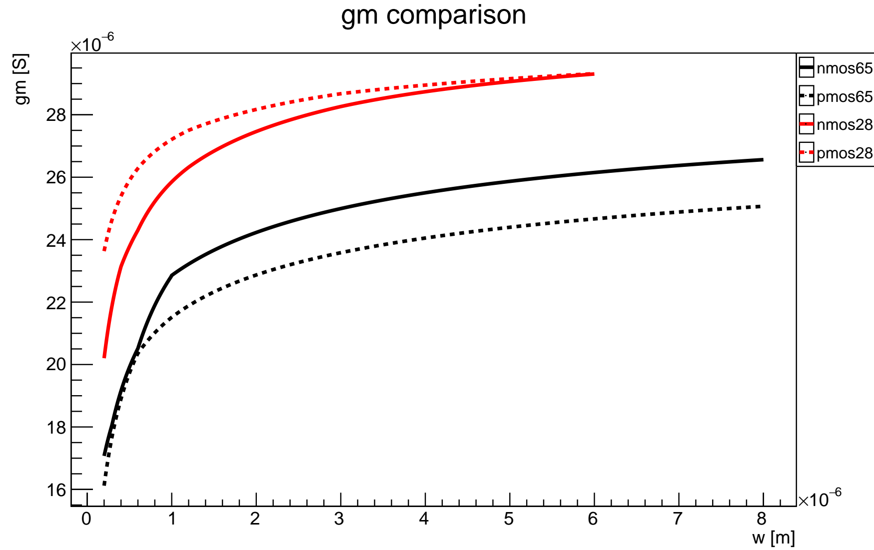


Figure 4.2: $g_m(W)$ of *NMOS* and *PMOS* transistors in 65nm and 28 nm evaluated with the configuration illustrated in fig.4.1, the $L = 100nm$ and $I_{ds} = 1\mu A$ values are fixed. The 28 nm models exhibits an overall larger g_m with the same sizing, this could be due to a large C_{ox} value as cause on the adoption of a HK dielectric. This is not a given parameter since the t_{ox} values of this two technologies are unknown. Other novel aspect is in the fact that for large W values the *PMOS* and *NMOS* transconductances are equalized: this is an effect of the adoption of strained channel material in *PMOS* which enhances the holes mobility. These curves were fitted using equation (2.23) with n and μC_{ox} as free parameters, the results of these fits are presented in tab.4.1.

$g_m(W)$ **fit parameters**		
	n	μC_{ox} $\left[\frac{\mu A}{V^2}\right]$
nmos 65	1.38	191
pmos 65	1.47	212
nmos 28	1.25	309
pmos 28	1.28	728

Table 4.1

Operating voltages

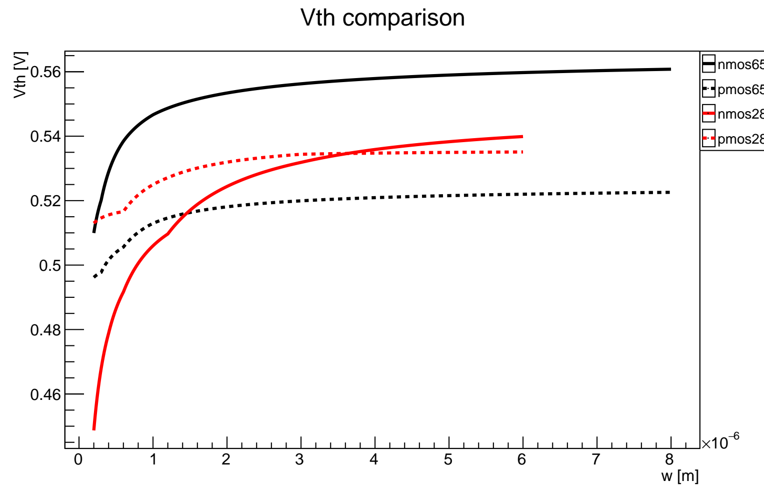


Figure 4.3: $V_{th}(W)$ of $NMOS$ and $PMOS$ transistors in 65nm and 28 nm evaluated with the configuration illustrated in fig.4.1, the $L = 100nm$ and $I_{ds} = 1\mu A$ values are fixed. The $PMOS$ curves represents $|V_{th}|$. The value of $PMOS$ and $NMOS$ V_{th} are equalized in 28 nm, this was possible thanks to the usage of different gate materials for the two different types of devices.

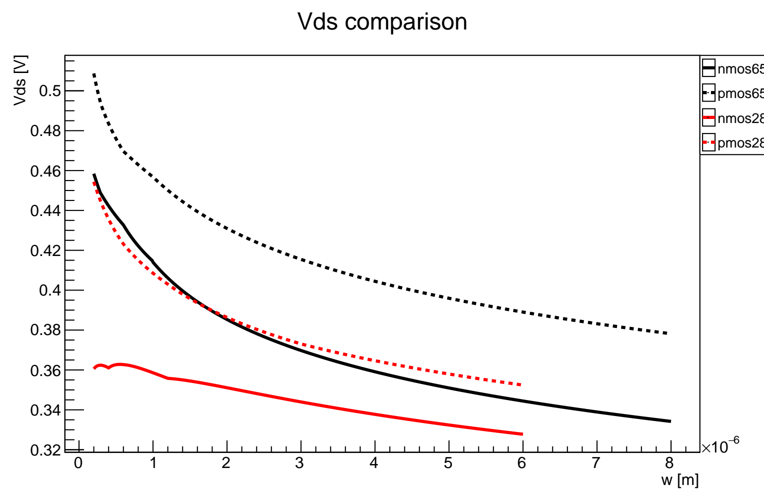


Figure 4.4: $g_m(W)$ of $NMOS$ and $PMOS$ transistors in 65nm and 28 nm evaluated with the configuration illustrated in fig.4.1, the $L = 100nm$ and $I_{ds} = 1\mu A$ values are fixed. In 28 nm less current is necessary to keep the transistors in saturation, this was probably done in order to ease the porting of designs from older technologies: both these voltages and V_{DD} were scaled of a factor ~ 0.83.

4.2 CSA Tests

The first analysis reported for the *CSA* is the transfer function of its core amplifier in order to verify the suitability of its performance in terms of the correct operation of the full system, this is done analysing its transfer function. A second analysis is on the shape of the full *CSA* signal. Both topics were discussed in the previous chapter.

4.2.1 Core Amplifier Transfer Function

The transfer function of the core amplifier where extracted from an AC simulation setted as the circuit in fig.4.5. The amplitude of the used sinusoidal voltage stimulus V_{ac} is $1\mu V$, a load capacitance $C_L = 1fF$ is used to emulate the capacitance of a potential input transistor of the next stage. The open loop gain A_0 and bandwidth BW of this stage are reported in tab.4.2 for both the 65nm and 28nm implementation.

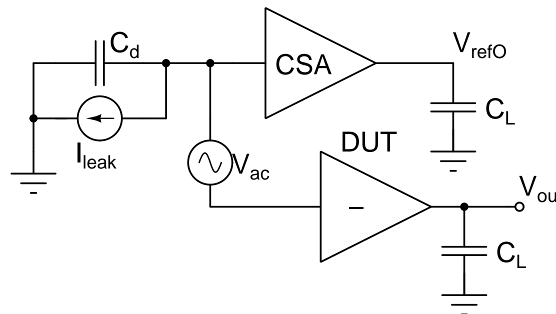
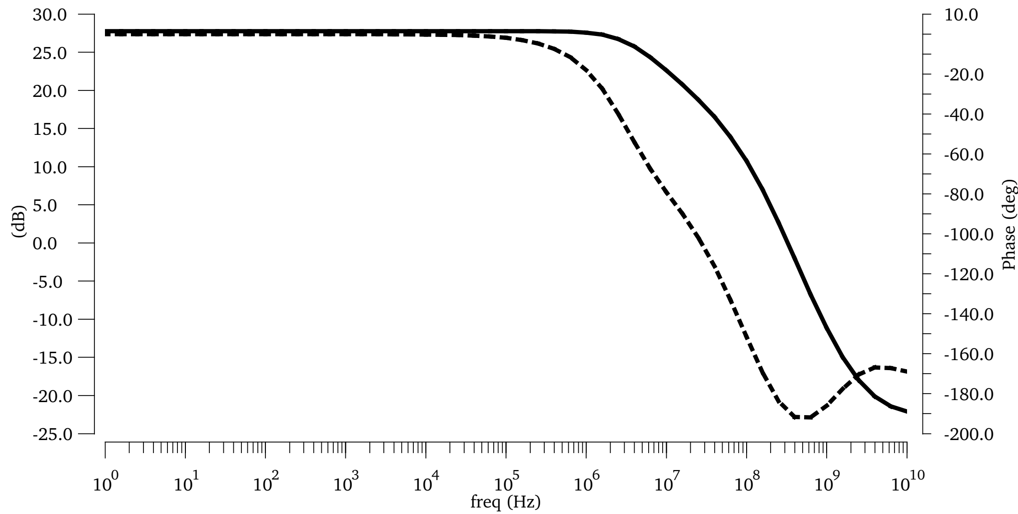


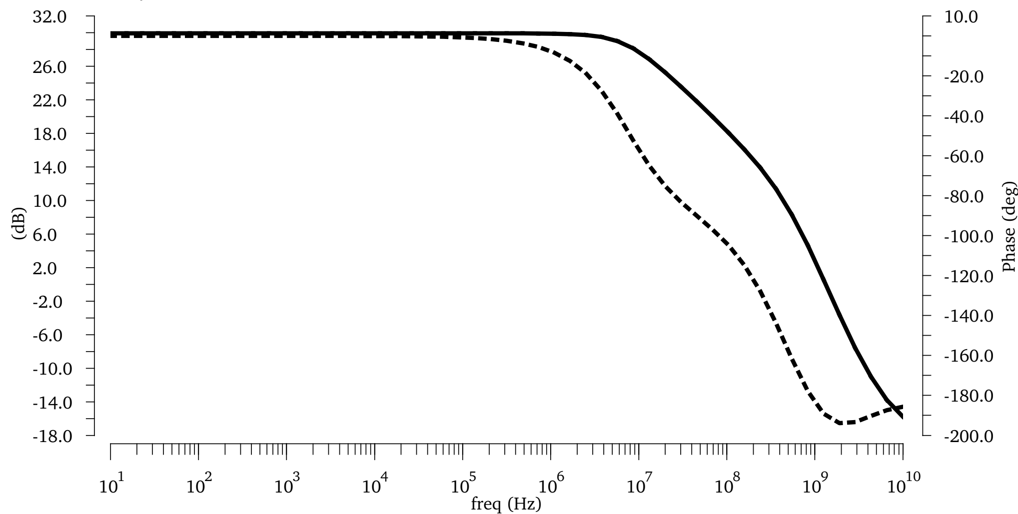
Figure 4.5: Circuit used to measure the transfer function of the core amplifier (indicated as device under test DUT). The voltage of the input node of the *CSA* is used as DC voltage point in order to obtain recreate the same condition of its target operation.

Core Amplifier Characteristics		
	A_0 $[dB]$	BW $[MHz]$
65 nm	27.7	5.9
28 nm	29.9	13

Table 4.2



(a) 65nm Core Amplifier Transfer Function: the gain is sufficient to make the *CSA* operate properly.



(b) 28nm Core Amplifier Transfer Function: the porting of the 65nm design of the cascode was carried out successfully maintaining its gain. The BW is also improved due to the reduction of $\frac{1}{4}$ of all transistors gate area.

Figure 4.6: The solid line is the gain of the amplifier, the dotted one is the phase. The transfer functions are obtained with the circuit in fig.4.5, the characteristics of the transfer functions are summarized in tab.4.2.

4.2.2 CSA Signal Characteristics

The response of the signal was studied with an actual current pulse extracted from a simulation of the silicon 3D sensor (conducted by INFN Cagliari). The pulse is presented in fig.4.7. The setup for this transient simulation is presented in fig.4.14, with a sensor capacitance $C_f = 100fF$, a load capacitance $C_L = 1fF$ and total current for the Krummenacher filter $I_k = 50nA$. The DC level was set to 170mV actin on V_{ref} of the *CSA*. The signals characteristics are summarized in tab.4.3.

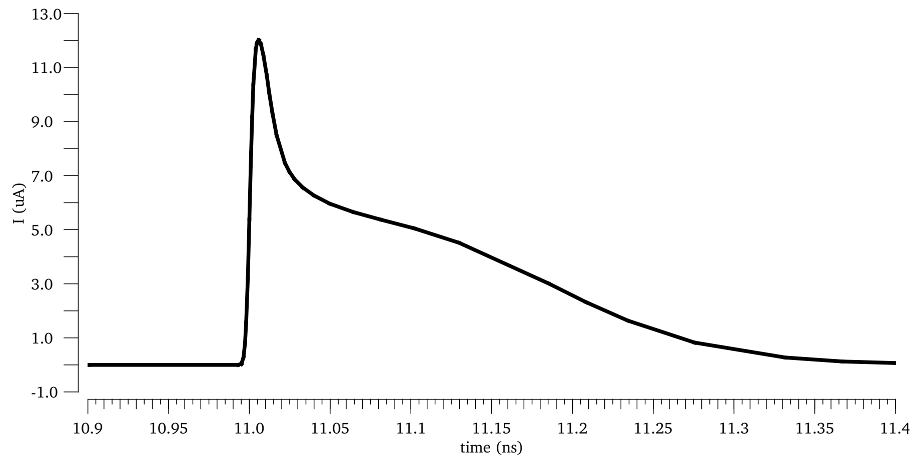
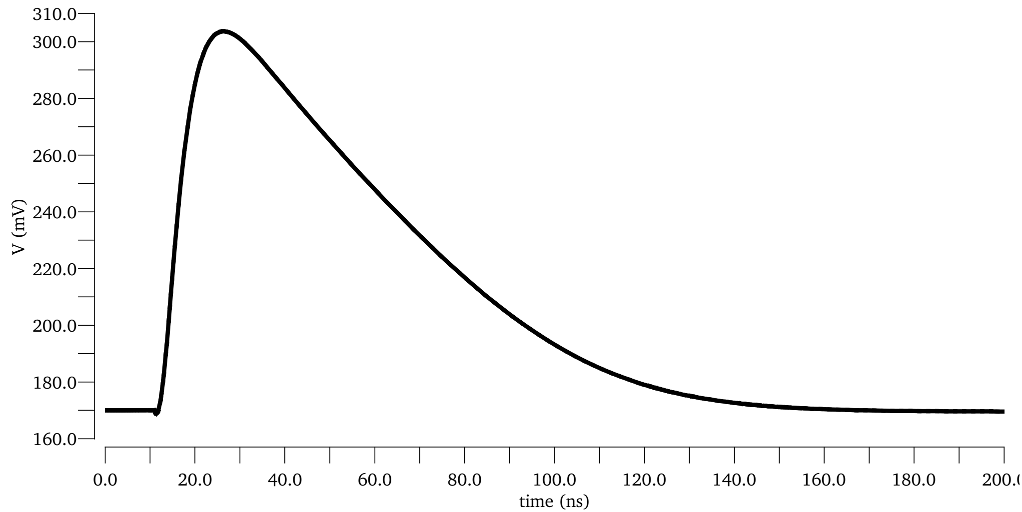


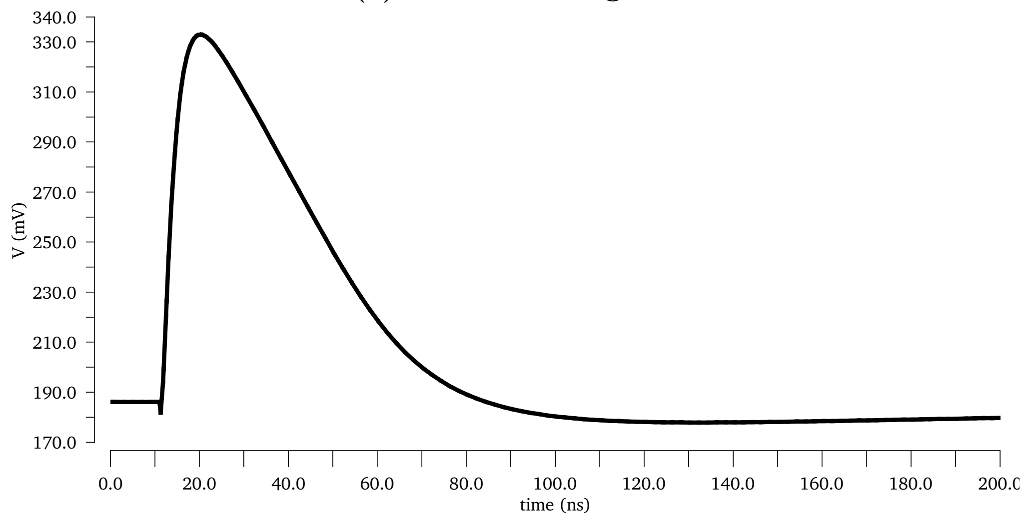
Figure 4.7: Simulated current pulse used as input signal for the *CSA*. This signal was extracted for a Ramo theorem based simulator. It is related to a $1fC$ charge, it has $12\mu A$ peak current and has a duration of $\sim 350ps$.

CSA Signal Characteristics		
	$G_c\ [\frac{mV}{fC}]$	$T_{pk}\ [ns]$
65 nm	107	16.3
28 nm	121	9.5

Table 4.3



(a) 65nm CSA Signal



(b) 28 CSA Signal: the 28nm version of the *CSA* features faster rising and falling times. This is probably due to the fact that both parameters depend on the inverse of a g_m (details in 3.2.2) since, as illustrated in 4.1, this parameter results larger in this technology for the same $\left(\frac{W}{L}\right)$.

Figure 4.8: Pulses obtained from a transient simulation in which a simulate signal of the actual sensor is used as input. The signals features the expected shape, even though its rise time is $\sim 3ns$ grater than the expected one . This is due to the core amplifier BW limitation. The signals characteristics are summarized in tab.4.3.

4.3 Leading Edge discriminator Tests

In this section the amplification stages of the leading edge discriminator are studied. This simulations were performed only on the 65 nm design since this circuit has not been ported to the 28nm node yet. The simulation were performed with the setup illustrated in fig.4.9. Two types of analysis were performed: the transfer function of the various stages in differential mode and the total transfer function for the common mode. The results of this analysis are summarized in tab.4.4.

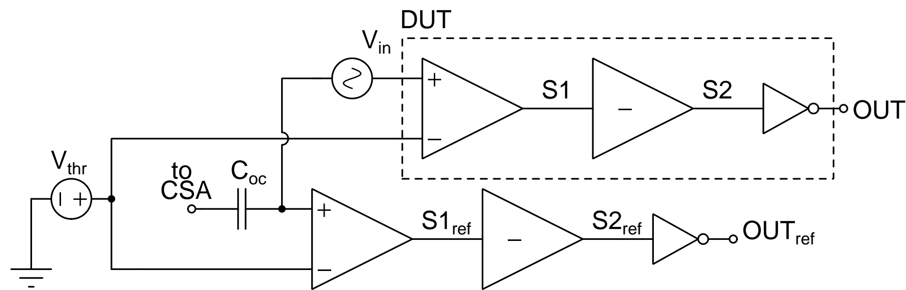


Figure 4.9: Setup used for the simulation of the discriminator stages transfer functions. The studied discriminator is a version of the final one without the offset correction circuit (DUT). In order to provide the properly off-setted inputs, the AC signal V_{in} was added on top of the base line DC level provided by an offset-corrected version of the discriminator. The DC level of this reference discriminator were also used to calculate the AC gain of the internal stages.

Discriminator Stages Characteristics		
	A_0 $[dB]$	BW $[MHz]$
S_1	12.1	56.3
S_2	17.0	13.1
total	29.1	13.3

Table 4.4

First Stage Transfer Function

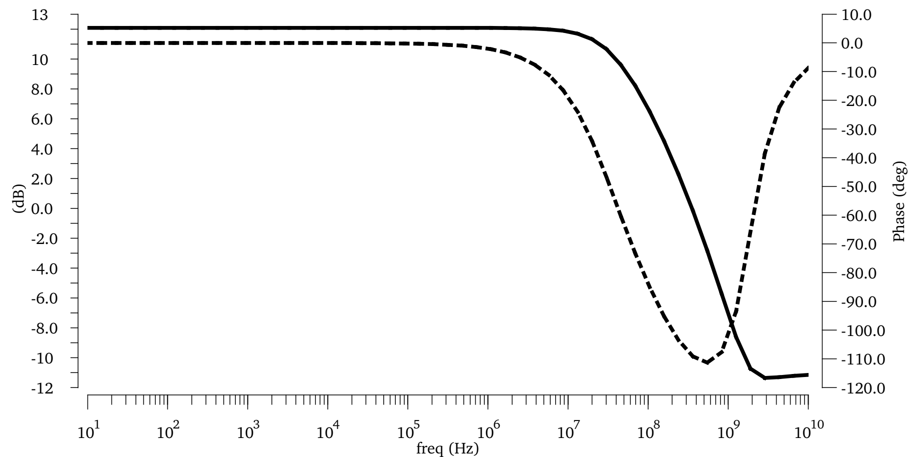


Figure 4.10: The solid line represents the gain, the dotted one is the phase. As discussed in 3.3.1 this low gain stage is used only to probe the differential mode between the signal and the threshold.

Second Stage Transfer Function

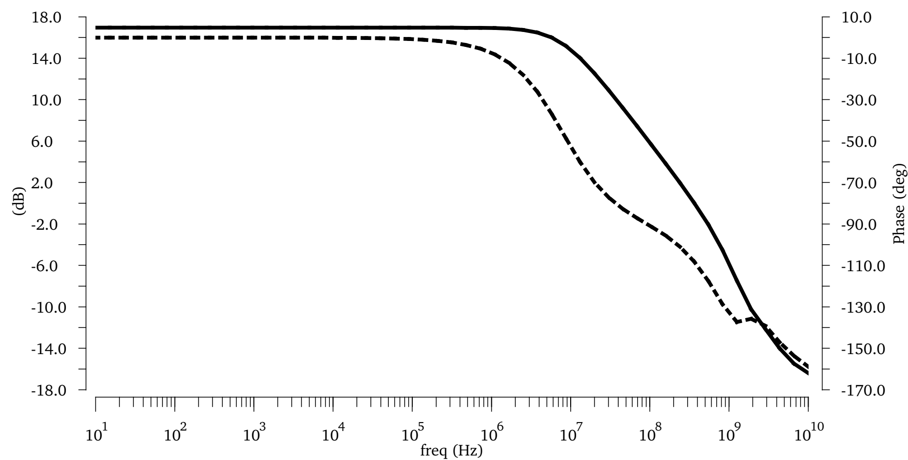


Figure 4.11: The solid line represents the gain, the dotted one is the phase. This stage constitutes the one the major gain: in spite of having the same input transistor sizing of the previous one, it exhibits a larger gain due to its cascode topology. The lesser BW of this is stage is probably due to the connection to the large output inverter.

Total Transfer Function

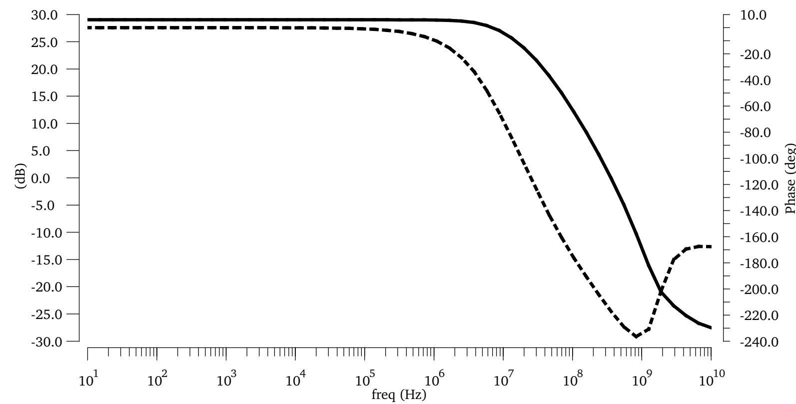


Figure 4.12: The solid line represents the gain, the dotted one is the phase. The total gain of the discriminator reaches almost 30dB, a good result considering the low power consumption and the fact that this is only a two stage topology.

Common Mode Transfer Function

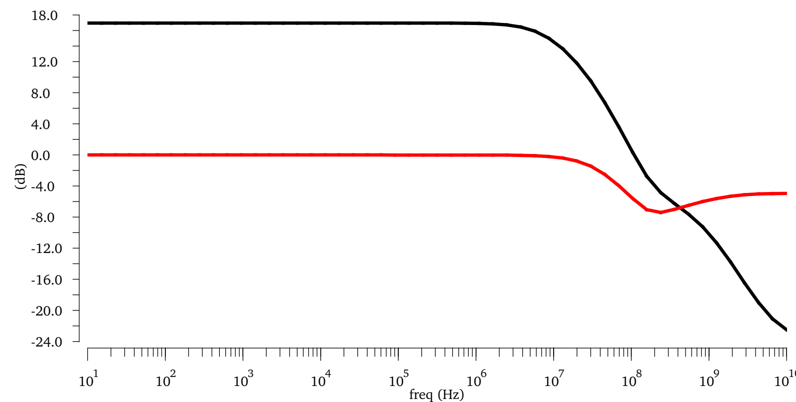


Figure 4.13: The black line is the transfer function of the whole amplification stage while the red is the one of the first stage. Most of the common mode gain is given by the second stage, while the first one slightly attenuates it. This is due to the fact that with the offset-correction the second stage is always in its most sensitive input range. The circuit can be destabilized by common DC changes, but the offset-correction mechanism will counteract this effect compensating the input. The input capacitor also works as an high-pass filter for the *CSA* terminal. The CMMR results 50.

4.4 Timing Tests

The following tests were performed in order to give a first evaluation of the timing precision of the system. All this results are extracted from transient simulations on the schematic circuit illustrated in fig.4.14. The signal characteristics were previously discussed in 4.2.2, the threshold level was set to $V_{thr} = 10mV$ on top of a $V_{bl} = 200mV$ base line. This value for the bas-line was chosen in order to both guarantee a large voltage range for the *CSA* signals and to let the discriminator operate in its most sensitive input range, in which the gain illustrated in the previous section could be met. As in the previous tests the voltage DC level at the output of the *CSA* is set to 170mV. The sensor capacitance was assumed $C_s = 100fF$ and the signals used as input are taken from a set of current pulses with $Q_{in} = 1.2fC$ charge, illustrated in fig.4.17.

The featured analysis ranges from the one of the jitter generated by the noise introduced by the front-end, the effect of the current signals shape variations, mismatch variations on the four transistors forming the differential pair of the discriminator, the effect of time-walk and the one on the process variation on the whole system. Most of this test were performed on the preliminary 65 nm design, only the noise simulation is performed on the 28nm design for now since the discriminator 28nm model is note yet completed.

Simulation Setup

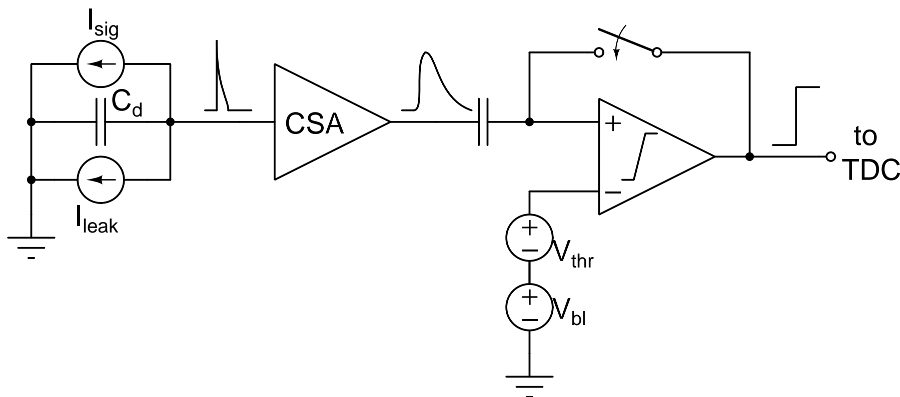


Figure 4.14: The simulated circuit includes the sensor model (on the left) with its capacitance, the *CSA* and the discriminator. The circuit is first set-up with the offset correction of the discriminator on the desired baseline and then the signal is presented.

4.4.1 Noise Contribution

The jitter is extracted from repeated noise simulations: in each simulation the noise contribution sources from each transistor (with its voltage and frequency distributions) are added to the deisgned schematic. What differentiate each simulation is in the usage of different seeds for the random number generator. In this analysis the contribution from the noise of both implementation of the *CSA* is presented.

65nm Signal with Noise

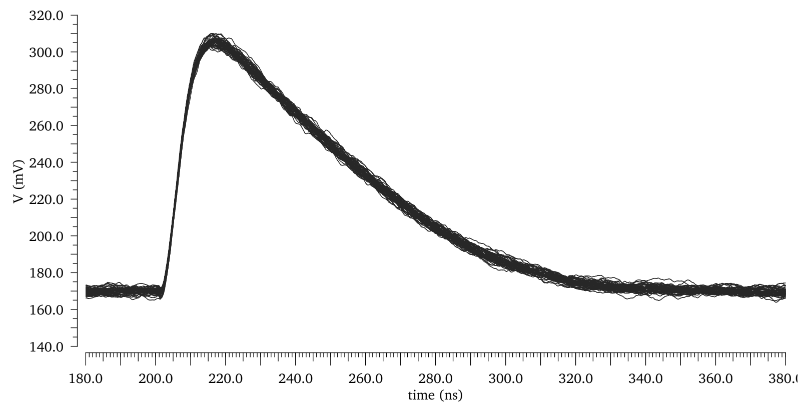


Figure 4.15: The plot presents 100 generated noise signals, the related SNR is $\sim$ 18.5 which result in a peak-to-peak jitter of 580ps (the estimated on is 676 ps).

28nm Signal with Noise

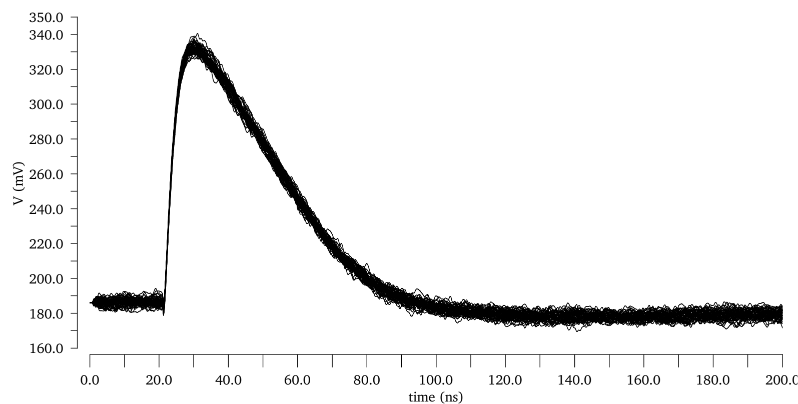


Figure 4.16: The plot presents 100 generated noise signals, the related SNR is $\sim$ 14.5 which result in a peak-to-peak jitter of 527ps (the estimated on is 633 ps). This contribution can be improved by reducing the peaking time or optimize the size of the most noise sensitive transistors of the design.

4.4.2 Signal Variation Contribution

The contribution on crossing time variation due to signal input variability was evaluated using the set of 5 signal presented in fig.4.17. This doesn't represents a statistical complete set, therefore a worst case scenario is considered. The analysis was conducted in the 65 nm version of the design. The impact of this effect on the output discriminator signal timing is $\sim 34ps$ peak-to-peak, this represents a good result in terms of the goal of a total variation $\Delta t_o < 100ps$ RMS. This optimal result is probably because at the small signal duration compared to the peaking time (350ps versus 12.5ns) which caused a good current integration.

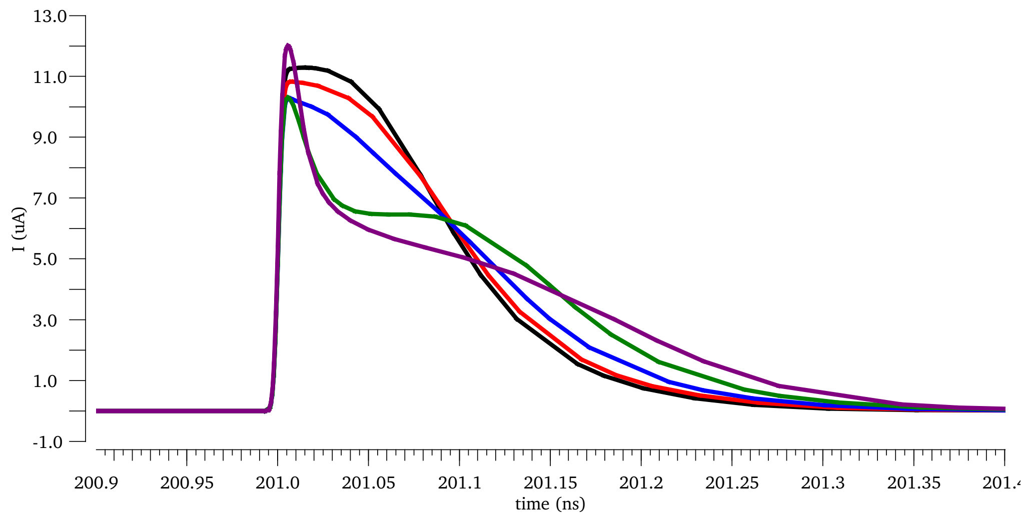


Figure 4.17: The set of five different current signals processed by the simulated front-end. The related charge Q_{in} is the same for every pulse is $1.2fC$, the shape variation is due to different impact point for a perpendicular to the top face incident particle. The charge development of the charge inside the material is linear and uniform. The signals with shorter peak amplitudes and longer sustain time are due to the presence of low filed zone inside the material.

4.4.3 Mismatch Contributions

The mismatch contribution has been simulated, as first test, on the differential pair of the discriminator since this constitutes the most mismatch sensitive topology in the front-end. Analysis of the timing variations of the discriminator delay t_d with and without offset correction were carried out.

Discriminator Mismatch contribution with Offset-Compensation

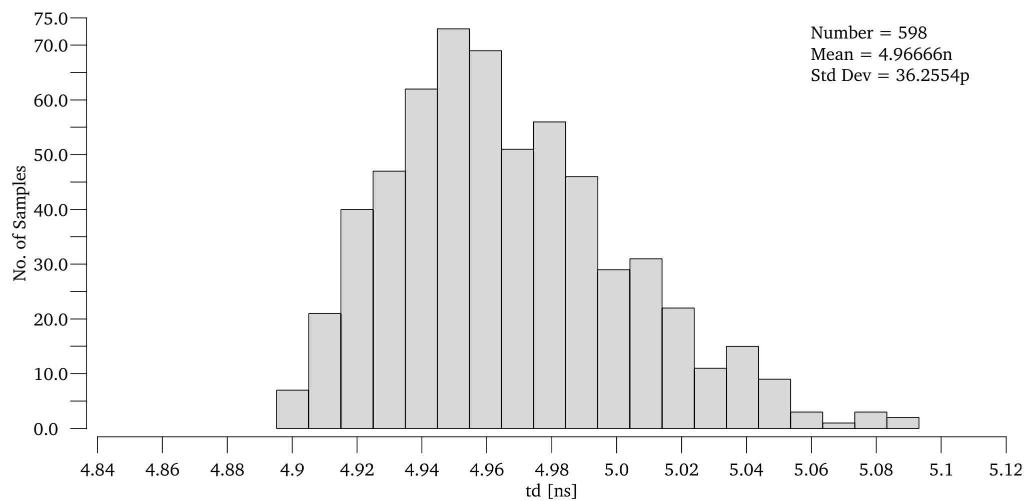


Figure 4.18: Distribution of discriminator delay time t_d due to mismatch effect. The standard deviation of this distribution results $\sigma_{mm} = 36ps$, an optimal result for the considered front end. A small population of high latency result can be observed, this is probably due the presence of unwanted mechanisms that can be probably eliminated through an optimization on the input transistors size.

Discriminator Mismatch contribution without Offset-Compensation

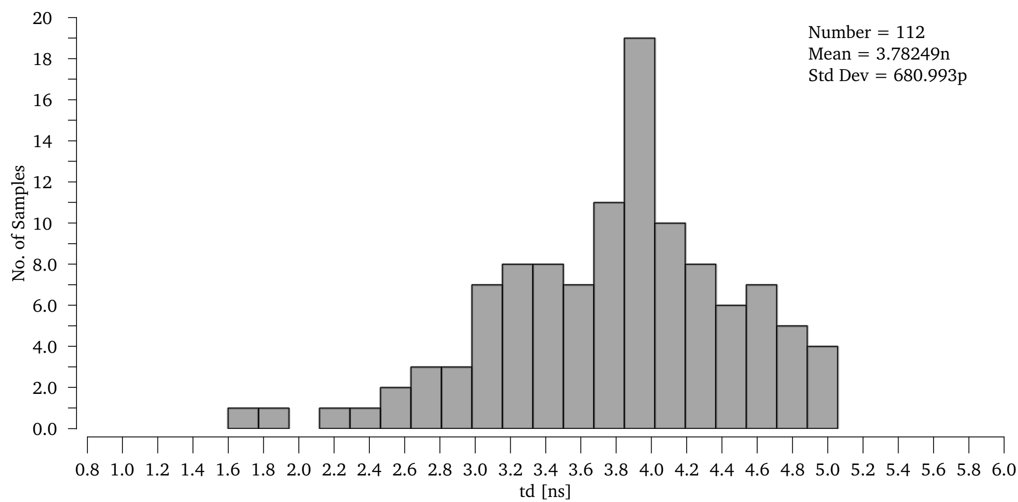


Figure 4.19: Distribution of discriminator delay time t_d due to mismatch effect. The standard deviation of this distribution results $\sigma_{mm} = 680ps$: more than an order of magnitude was gained in this aspect introducing the offset-correction circuit. It must be noted that some of the runs have failed: in some case the mismatch contribution was so large that made the DC level at the output of the first stage shift to low voltages, in this way the output state of the discriminator was locked at logic 1.

4.4.4 Time Walk

The time-walk measure was performed assuming, in the absence of actual sensor simulation, a constant time duration for different total charges developed inside the sensor. In this way the signal used to carry out the simulation was stretched-out on the current direction, an example of resulting signals is presented in fig.4.20. The resulting voltage signals are presented in fig.4.21. Three parameter were extracted from this simulation:

- t_c: the ideal threshold crossing time of the *CSA* signal relative to the start of the rising edge of the current signal, this term accounts only for the discriminator contribution to time-walk.
- t_d: the delay time between t_c and the point in which the digital signals reaches 400mV, this is the contribution to the time-walk added by the discriminator.
- $t_o = t_c + t_d$: the total delay time elapsed between the arrival of the current signal and the digital response.

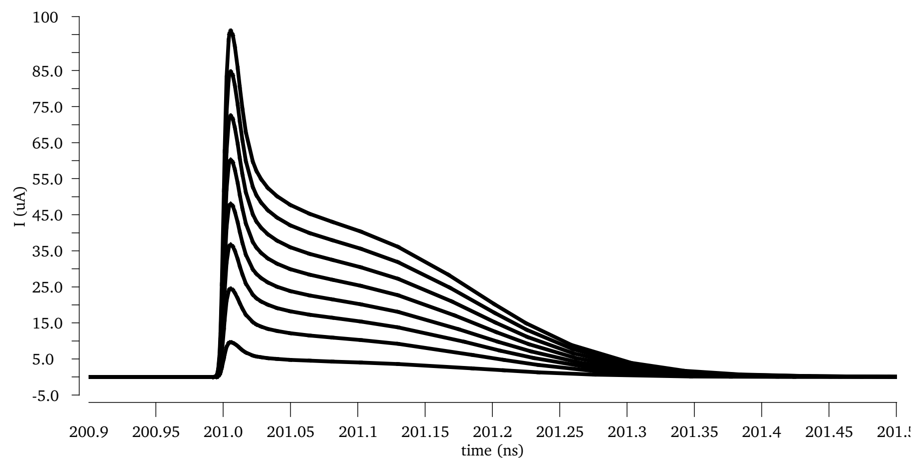


Figure 4.20: An example of current signals used for time walk evaluation obtained by stretching the pulses in the current direction. Their related charges ranges from $1fC$ to $10fC$

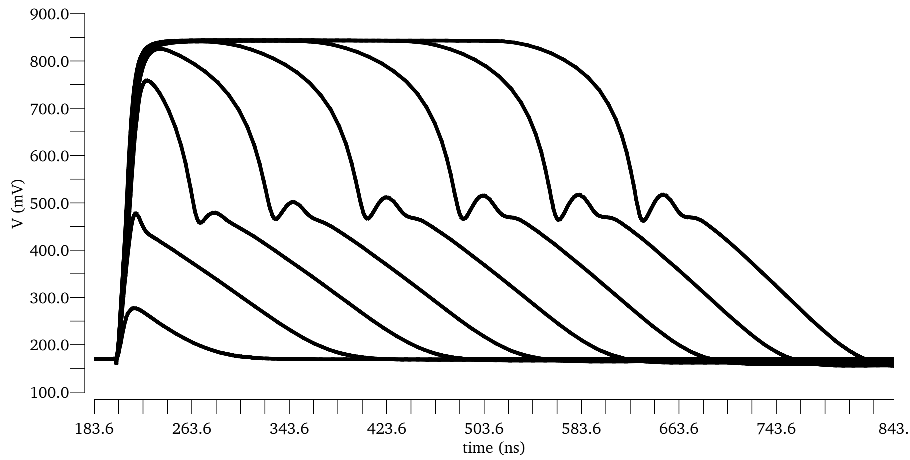


Figure 4.21: An example of the voltage pulses obtained by processing the signals of fig.4.20.

Peaks Amplitude

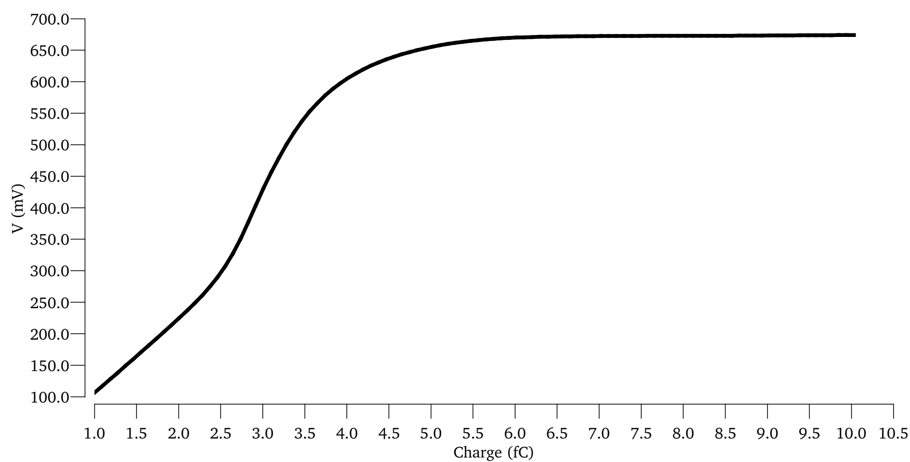


Figure 4.22: Trend of the peaks amplitude in relation to the input charge: the trend is almost linear for charges lower than fC, after this point the CSA saturates its input range with major distortions over $4fC$.

CSA time-walk contribution (t_c)

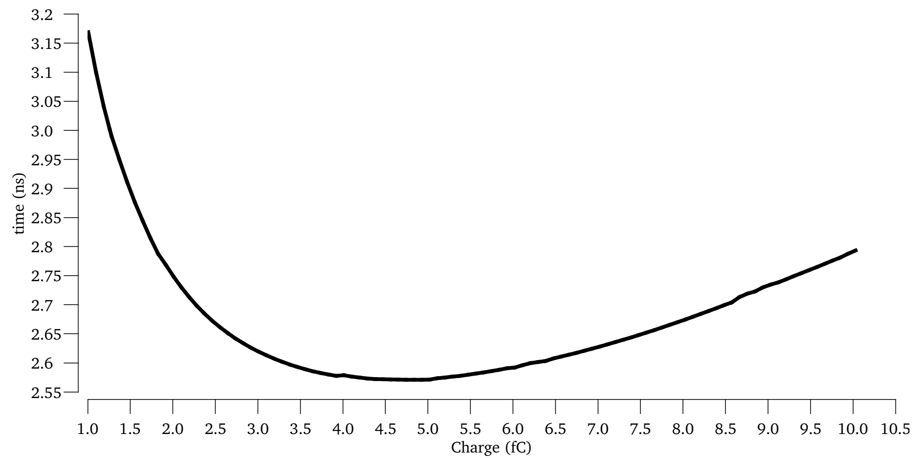


Figure 4.23: Ideal threshold crossing time of the *CSA* signals for different input charges. The largest time difference is 598 ps, this fact suggests that a simple leading edge discriminator approach is no suited for this type of signals. The trend is also non-monotonic, this issue needs to be correct since it doesn't allows *ToT* correction.

Discriminator time-walk contribution (t_d)

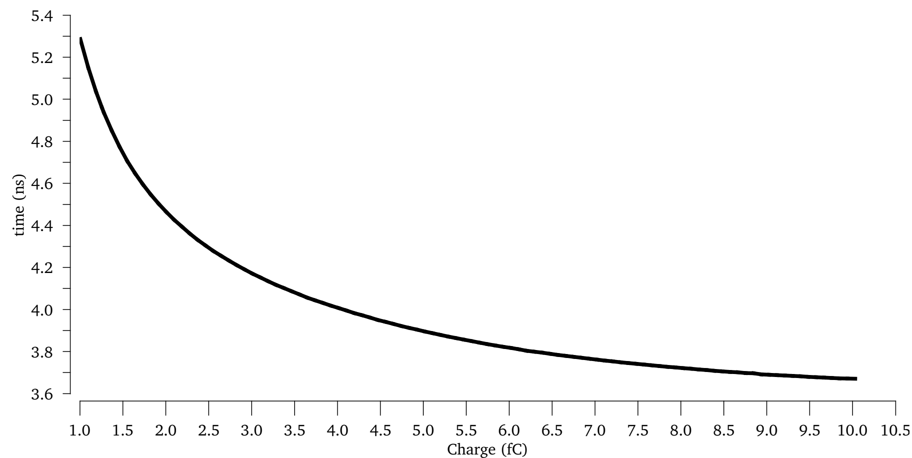


Figure 4.24: Variations on discriminator delay time in relation to different input charges. The largest time difference is 1.62 ns: the discriminator introduces the major contribution to the time-walk. This could be cause by a residual slope sensitivity of the discriminator which can be improved increasing its gain. The monotonic behaviour of this contribution makes *ToT* correction viable.

Total time-walk of the front-end (t_o)

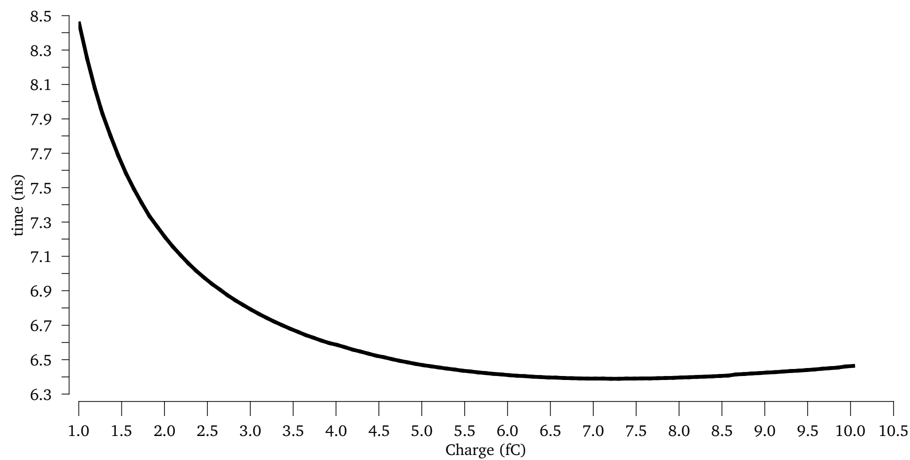


Figure 4.25: Total time-walk of the system: the total time walk of the system is 2.97 ns for input charges that ranges between $1fC$ and $10fC$. The system requires to take in account this effect: CFD approach can be explored, while at the current state of the design a ToT correction is not possible due to the overall lack of monotonicity of this trend.

Ideal *ToT* measure

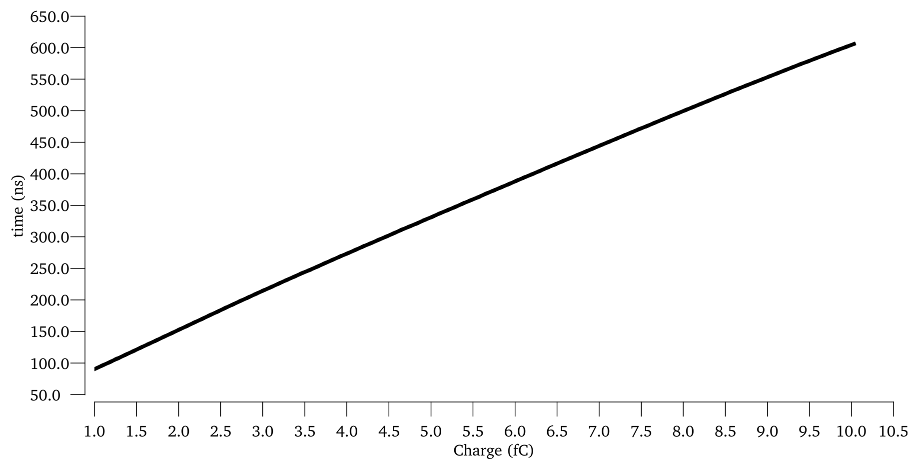


Figure 4.26: Ideal Time-over-Threshold measure of the *CSA* signals in relation to the input charge: it exhibits a good linear trend which suggests the possibility of ToT correction since the total time-walk is almost monotonic.

4.4.5 Process Variations Contributions

An analysis on the process variations contribution to the timing variability of the signal is presented. These consists in long range variations of the characteristics of the devices in the wafer, for this reason the impact of channel-to-channel variations will be small in a small area chip (compared to the total wafer size). Process variations can be corrected chip-to-chip with dedicated calibration circuits.

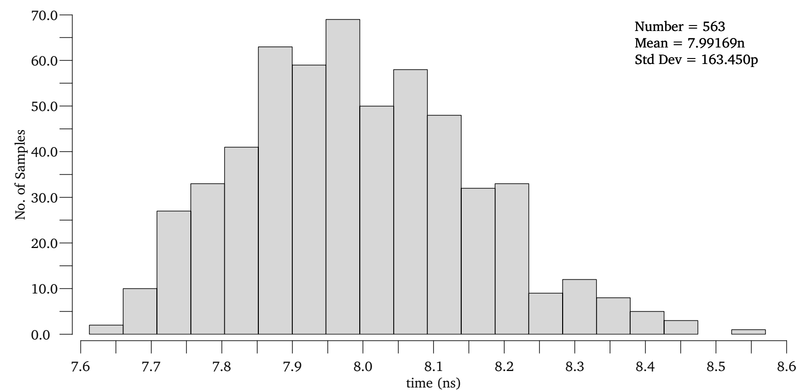


Figure 4.27: Effect of process variations on the total time of the discrimination t_o, as can bee seen the variations are rather small with an standard deviation of 163ps. This result is obtained thanks to the offset correction mechanism since the largest variation is in the DC levels of the circuit.

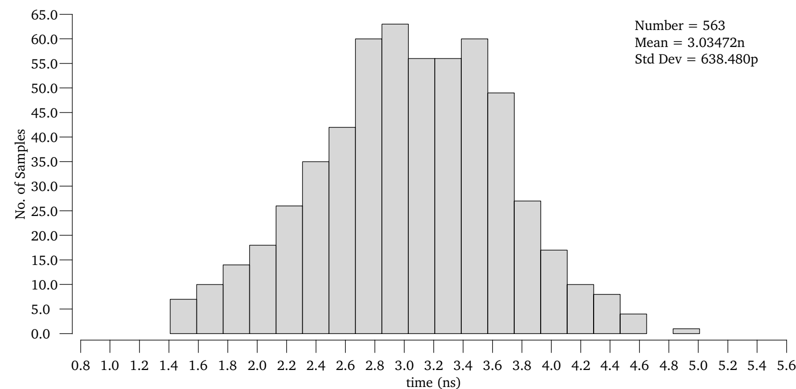


Figure 4.28: Effect of process variations on threshold crossing of the *CSA* signals t_c. This variation is fairly large (638ps standard deviation) due to the DC base-line variation. This contribution is well canceled by the input base line setting operated by the offset-correction circuit.

This work was carried out as part of the INFN TIMESPOT project, which aims to research and produce a working prototype tracking detector of the ones that will be used in future High Luminosity HEP experiments such as *HL-LHC*. The requirements of this type of experiments are high time and space resolution, high even-rate and elevate radiation hardness. This has required an interplay between the various INFN research centers involved in the project. By the front-end point of view, the innovation introduced by the adoption of currently in development 3D sensors has opened the possibility to co-design simultaneously these two parts of the detector. This has let both sides to define the characteristics demanded by the other part in order to meet the design goals, focusing the optimization of the single device performances on key aspects.

The design of this front-end was carried in all of its phases using high standard level tools and techniques for circuit implementation ,verification and optimization. Simulations results indicates that the proposed front-end architecture operates properly under constrain of chip area and power consumption dictated by necessities of pixel-detectors electronics. This consists firstly in a low power consumption charge sensitive amplifier which is able to operate properly with the strict conditions imposed by sensor characteristics. The second part of the front end consists in a first tentative discriminator designed to obtain high reliability in timing signal generation. The feasibility of the proposed design was studied using industry leading simulation techniques. The input amplifier architecture was successfully ported from a standard 65nm node to the novel 28 nm one, with a general improvement on its performance.

Timing performance of the design shows promising results in terms of input signal variation, mismatch and process variations. These results are mainly due to key aspect integrated in the proposed design such as DC current compensation and voltage offset compensation techniques. As for noise induced fluctuation on the signals timing, the simulation revealed that further optimization steps are required in order to meet the desired goals.

This optimization phase as well as the design of the physical layout of the circuit represents the first natural progression on the presented work. The next step would be to port the discriminator design in the 28 nm technology as well build-up on

this core block to design an implementation of a more sophisticated discrimination approach in order to correct the time-walk effect. Further development will range in other areas of interests on this application such as the design of the time to digital converter which will be coupled with the proposed analog front-end, as well as the integration on the final system on chip.

In the following months a first test chip, which will include the analog block designed in this work, will be produced. Several operation tests on this prototype, including radiation hardness, are planned for the near future.

Bibliography

[1] G. Apollinari, O. Bruening, T. Nakamoto, and L. Rossi. High luminosity large hadron collider hl-lhc. *CERN Yellow Report*, 5 2015.

[2] Angelo Rivetti. New needs and directions in microelectronics and ultrafast electronics. *Proceeding of Science*, 12 2015.

[3] The timespot project. INF CSNV Open Call, 2017.

[4] Helmuth Spieler. *Semiconductor Detector System*. Oxford University Press, Physics Division, Lawrence Berkeley National Laboratory, 2005.

[5] Gian-Franco Dalla Betta. 3d silicon detectors. *Proceeding of Science*, 3 2014.

[6] S. Lagomarsino et al. Three-dimensional diamond detectors: Charge collection efficiency of graphitic electrodes. *Applied Physics Letters*, 103, 12 2013.

[7] G. Parrini et al. Laser graphitization for polarization of diamond sensors. *Proceeding of Science*, 7 2011.

[8] Angelo Rivetti. *CMOS Front-End Electronics for Radiation Sensors*. CRC Press, INFN Torino, Italy, 2015.

[9] Sorin Martoiu et al. A low power front-end prototype for silicon pixel detectors with 1oops time resolution. *IEEE Nuclear Science Symposium Conference Record*, 11 2008.

[10] Yannis Tsividis. *Operation and Modeling of The MOS Transistor*. Oxford University Press, Columbia University, 2 edition, 2003.

[11] H. Nyquist. Thermal agitation of electric charge in conductors. *Physical Review*, 32, 7 1928.

[12] Jung-Suk Goo et al. Physical origin of the excess thermal noise in short channel mosfets. *IEEE Electron Device Letters*, 22(2), 2 2001.

[13] Marcel J. M. Pelgrom, Aad C. J. Duinmaijer, and Anton P. G. Welbers. Matching properties of mos transistors. *IEEE Journal of Solid-State Circuits*, 24(5), 10 1989.

[14] K. Mistry et al. 45nm logic technology with high-k+metal gate transistors, strained silicon, 9 cu interconnect layers, 193nm dry patterning, and 100% pb-free packaging. *IEEE*, 12 2007.

[15] Kelin K Khun. Cmos scaling beyond 32nm: Challenges and opportunities. *IEEE*, 7 2009.

[16] John Robertson. Band offsets and work function control in field effect transistors. *J. Vac. Sci. Technol. B*, 27(1), 2 2009.

[17] S. Deora et al. Intrinsic reliability improvement in biaxially strained sige p-mosfets. *IEEE EDL*, 32(3), 3 2011.

[18] K. Rim et al. Characteristics and device design of sub-100 nm strained si n- and pmosfets. *IEEE Symposium On VLSI Technology Digest of Technical Papers*, 2002.

[19] T. Ghani et al. A 90nm high volume manufacturing logic technology featuring novel 45nm gate length strained silicon cmos transistors. *IEEE IEDM*, 12 2003.

[20] Frank Schellenberg. A little light magic. *IEEE Spectrum*, 9 2003.

[21] Marc D. Levenson, N.S. Viswanathaan, and Rovert A. Simpson. Improving resolution in photolithography with a phase-shifting mask. *IEEE Transactions on Electron Devices*, 29(12), 12 1982.

[22] A. J. Annema, B. Nauta, R. Lengevelde, and H. Tiunhout. Analog circuits in ultra-deep-submicron cmos. *IEEE Journal of solid state circuits*, 40(1), 1 2005.

[23] Wei Zhao anff Yu Cao. New generation of predictive technology model for sub-45 nm early design exploration. *IEEE Transaction on electron devices*, 53(11), 11 2003.

[24] Srinivas Raghvendra and Philippe Hurat. Dfm: Linking design and manufacturing. *IEEE VLSID'05*, 2005.

[25] Jörn Lange. Recent progress on 3d silicon detectors. *Proceeding of Science*, 11 2015.