Università degli Studi di Torino
Scuola di dottorato in Scienze della Natura e Tecnologie Innovative
Dottorato in Fisica e Astrofisica

Front-end electronics in 65nm CMOS technology for the HL-LHC upgrades

Candidate: **Ennio Monteil**

Supervisor: **Dott. Angelo Rivetti**

XXIX ciclo

Anni accademici: 2014-2015-2016

Settore Scientifico Disciplinare FIS/01

Abstract

After 2025 the operation of the High Luminosity Large Hadron Collider (HL-LHC) will be started. The luminosity of the machine will be increased from $1 \times 10^{34} cm^{-2} s^{-1}$ to around $7 \times 10^{34} cm^{-2} s^{-1}$. Based on the important physics results obtained at LHC after 2009, the purpose of the upgraded version of the machine is to allow further inspection of the Higgs and Standard Model (SM) physics, together with the search of new channels beyond the SM. The experiments, like the Compact Muon Solenoid (CMS), are required to undergo an upgrade program, the so-called Phase 2, to be able to fully profit of the increased performance of LHC. From the point of view of the inner silicon pixel tracking system, in fact, unprecedented levels of radiation (up to 1 Grad in 10 years), hit rate (3 GHz/cm^2 for the innermost layer) and event Pile-Up (140-200 collisions per bunch crossing) will be reached. Since the current implementation of the CMS inner tracking system would not be able to cope with these extreme requirements, the development of a new silicon pixel detector is needed. It will be composed of hybrid pixel detectors, in which the silicon sensor is bump-bonded to the readout chip. The size of the single pixel will be reduced to 50×50 μm^2 or 25×100 μm^2 in order to manage the occupancy due to the high rate. For the design of the readout chip, the RD53 Collaboration has been established at CERN in 2013, with a strong involvement by INFN through the CHIPIX65 project, funded by the INFN R&D committee. A CMOS 65 nm technology has been chosen for this ourpose, given that it combines a high radiation tolerance and the possibility of implementing a complex digital intelligence in the small area available. The Ph. D. activity presented in this work has been carried out in this framework and has been focused on the development of an analog front-end chain for the readout chip of the Phase 2 CMS silicon pixel detector.

The design of dedicated chips for the readout of silicon sensors In High Energy Physics (HEP) has started at the beginning of the 1980s. From the point of view of the analog front-end, the first designs featured discrete-time architectures based on switched capacitor techniques. Since this kind of solution requires the implementation of a number of switches in each channel, the following analog front-ends for HEP applications have been characterized by continuous-time designs. In fact, at that time the large size of the transistors, beyond 1 μm, led to important charge injections during the opening and closing transitions, generating also additional noise contributions. Nevertheless, the implementation of discrete-time architectures has been carried on in other applications like ADCs, leading to significant improvements in the most critical aspects. Nowadays, thanks to the huge scaling of CMOS devices, the choice of discrete-time front-ends becomes competitive again also for HEP application. In fact, the choice of a CMOS 65 nm technologies allows to keep charge injection effects and the related consequences on noise under control. In this work, a topology in which a continuous-time input stage is coupled to a discrete-time discriminator has been chosen, with the purpose of optimizing the noise performance and the speed-power trade-off.

The analog front-end is organized in two main building blocks: the preamplifier and the discriminator. Given the small size of the signal generated by the sensor, the former has the goal of providing a large gain to allow a proper signal processing. As a consequence, a telescopic cascode architecture has been chosen. The relative feedback network is based on the well-

known Krummenacher circuit. It has been used to provide a constant current discharge of the feedback capacitance. This feature is in fact crucial for the storage of the Time-over-Threshold information, which is proportional to the charge deposited in the sensor. The discriminator contains the most innovative solutions from the point of view of the analog front-ends for the readout of silicon sensors. It is a synchronous stage: the preamplifier signal is sampled only at the rising edge of the LHC 40 MHz clock. This choice is motivated by the fact that in the experiments collisions take place only every 25 ns, making feasible the use of a discrete-time discriminator. In addition, a switched capacitor-based solution has been adopted to address a common issue for analog front-end chains: the compensation of the offset in the differential amplifiers, implemented to provide a small additional gain and the threshold voltage setting. In this fornt-end, in fact, the offset is stored on the capacitors during a short calibration period. In comparison to the usual compensation performed by using a local DAC, this technique offers the advantage that no additional trimming procedure is required. The second part of the discriminator is composed of a positive feedback latch which performs the proper comparison between the signal and the threshold. With the purpose of up to 8-bit storage of the ToT information, this stage can also be automatically turned into a local fast oscillator with adjustable frequency. Such an approach has been used in some applications like SAR ADCs, but it is new in the field of the readout of silicon pixel sensors for HEP experiments.
All the building blocks, together with the results of CAD simulations and of measurements of test prototypes, are extensively discussed in this work, which is structured as follows:

- In **Chapter 1**, the CMS experiment and its upgrade programs are described in detail, together with an overview of the physics results obtained in the first year of operation and the perspectives for new discoveries in the next years;
- **Chapter 2** is focused on the application of CMOS technologies to analog design. The short channel effects, additional phenomena which characterize the deep submicron transistors, are discussed, together with the performance degradation induced by the radiation;
- In **Chapter 3**, the basic blocks of an analog front-end and their main features are listed;
- **Chapter 4** is dedicated to the description of the analog front-end studied during the Ph. D. The design choices of the building blocks (preamplifier, feedback network, discriminator) are discussed in detail, together with CAD simulations results;
- **Chapter 5** concerns instead the test results of two small prototypes submitted for production to the foundry;
- In **Chapter 6** the large demonstrator chips designed by CHIPIX65 and RD53A collaborations respectively are presented.

Contents

Chapter 1

CMS experiment results and motivations for the Phase 2 upgrade

This chapter provides an overview of the CMS experiment as it started operation in 2009, with a particular focus on the tracking detector, together with the main physics results already published. Subsequently, a detailed description of the Phase 2 upgrade of CMS foreseen for the High-Luminosity Large Hadron Collider (HL-LHC) is presented. A particular focus on the silicon pixel detector is given in this chapter, together with the main specifications to be considered in the analog front-end readout chip, in order to provide the background for the Ph. D. research activity described in this work.

1.1 The CMS experiment

CMS (Compact Muon Solenoid) [1] is one of the four experiments located at the Large Hadron Collider (LHC) [2] at CERN, along with ATLAS, ALICE and LHCb. It is a general purpose experiment, targeting the discovery of new particles and the study of a wide range of phenomena. LHC, a 27-km ring of superconducting magnets, has been designed for head-on collisions of 7 TeV proton beams at a luminosity of $10^{34}\ cm^{-2}s^{-1}$ and of ion beams with 2.75 TeV per nucleon at a luminosity of $10^{27}\ cm^{-2}s^{-1}$. These features allow the study of a wide range of physics phenomena [3]:

- Electroweak symmetry breaking caused by the Higgs mechanism and Standard Model
- Supersymmetric particles, such as squarks and gluinos
- New massive vector bosons, like Z'
- Extra dimensions
- Heavy-ions physics

As a consequence, the initial detector concept has been conceived to be able to explore all the physics channels presented in the above list. The main requirements were:

- Good muon identification and momentum resolution over a wide range of momenta in the region $|\eta| < 2.5$
- Possibility of resolving unambigously the charge of muons with $p < 1\ TeV/c$
- Good charged particle momentum resolution and reconstruction efficiency in the inner tracker

- Good electromagnetic energy resolution, good diphoton and dielectron mass resolution (1% at 100 GeV/c^2) and wide geometric coverage ($|\eta| < 2.5$)

The timeline of the LHC operation is presented in figure 1.1 [4]. The LHC machine has started its operation in 2009, producing 7-8 TeV center-of-mass collisions until the beginning of 2013. This first period of data-taking has been called Run 1. After the first Long Shutdown (LS1), the accelerator performance has been improved, reaching the nominal values in luminosity and in center-of-mass energy (Run 2). The Run 3 will feature an additional improvement in luminosity and will close the so-called Phase 1 of the experiments. Recently, a further upgrade of the collider, the HL-LHC, corresponding to the Phase 2 of the experiments, has been approved for a 10-year operation starting in 2026. The total integrated luminosity will be an order of magnitude higher compared to the standard LHC runs. As explained in detail in section 1.4, HL-LHC will put severe requirements on the detectors, leading to extended upgrade programs.

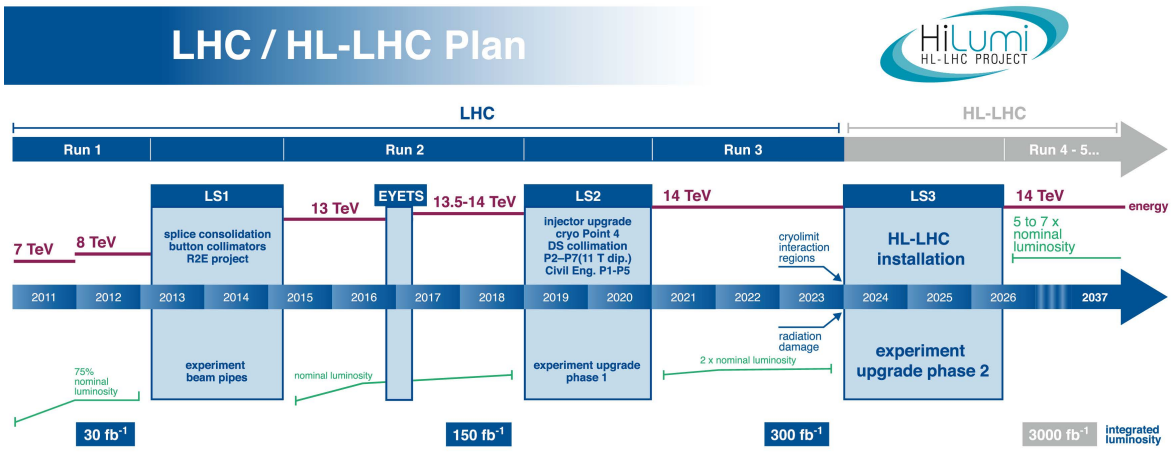

Figure 1.1: Timeline of LHC operation [4]

1.2 CMS first years of operation

In this section the layout of the CMS detector used since the beginning of data taking is presented, with a particular focus on the tracker detector. In addition, a quick overview of the main physics results collected so far during Run 1 is illustrated.

1.2.1 Detector overview

An important aspect driving the detector design and layout is the choice of the magnetic field configuration for the measurement of the muons momentum. Large bending power is needed to measure precisely the momentum of high-energy charged particles. This aspect forces a choice of superconducting technology for the magnets. [5]. A transverse section of the CMS detector, which is 21.6-m long and has a diameter of 14.6 m, is shown in Figure 1.2.
The main features are the high magnetic field solenoid, the fully silicon-based inner tracking system and a fully active scintillating crystals-based electromagnetic calorimeter.
The detector is build around the superconducting magnet, able to generate a magnetic field equal to 3.8 T. The bore of the magnet coil is large enough to accommodate the inner tracker and the calorimetry inside. The silicon tracker is placed close to the interaction region to improve the measurement of the impact parameter of charged-particle tracks, as well as the position

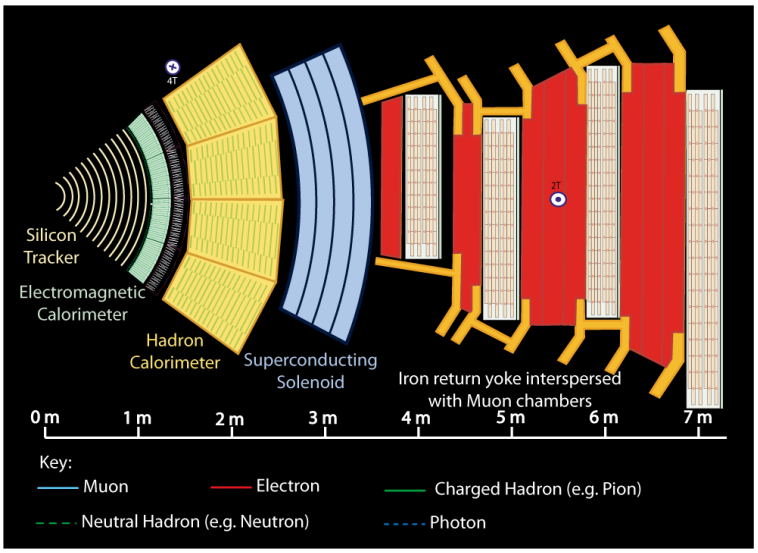

Figure 1.2: Transverse section of the CMS experiment

of secondary vertices. The electromagnetic calorimeter (ECAL) uses lead tungstate ($PbWO_4$) crystals with coverage in pseudorapidity up to $|\eta| < 3.0$. A preshower system is installed in front of the endcap ECAL for π_0 rejection. The ECAL is surrounded by a brass/scintillator sampling hadron calorimeter (HCAL) with coverage up to $|\eta| < 3.0$. The outer part consists of muon stations, each of them composed of several layers of aluminium drift tubes (DT) in the barrel region and cathode strip chambers (CSC) in the endcap region, complemented by resistive plate chambers (RPC) [5].

1.2.2 Tracker design and performance

The inner tracking system of CMS is designed to provide a precise and efficient measurement of the trajectories of charged particles emerging from the LHC collisions, as well as a precise reconstruction of secondary vertices, for a working period of about 10 years. The operating conditions of the experiment are particularly challenging. At the LHC design luminosity of $10^{34}\ cm^{-2}s^{-1}$ on average about 1000 particles from more than 20 overlapping proton-proton interactions cross the tracker for each bunch crossing (every 25 ns). As a consequence a detector technology featuring high granularity and fast response is required, such that the trajectories can be identified reliably and attributed to the correct bunch crossing. In turn, these aspects lead to the necessity of implementing a high power density on the electronics side, requiring also an efficient cooling system [5]. On the other hand, the material budget should be kept as low as possible in order to reduce effects like bremsstrahlung and multiple scattering, leading to the necessity of finding a trade-off between these two requirements. In addition, the high particle flux results in a severe demand about radiation effects. Bearing in mind all these aspects, it was decided to put in place a tracker fully based on silicon detector technology. The read-out chips employed in the CMS tracker are fabricated in standard 0.25 µm CMOS technology.
As shown in Figure 1.3, it is composed of two main parts: the silicon pixel detector and the silicon strip detector. The pixel detector consists of three concentric cylindrical barrel layers and four blade disks which close the barrel ends. The barrel layers have an active length of 53 cm and are located at average radii of 4.3, 7.3, and 10.2 cm. The endcap disks instrument the regions between radii 4.8 and 14.4 cm at mean longitudinal distances of 35.5 and 48.5 cm from the interaction point [6]. The system provides efficient three-hit coverage in the region of pseudorapidity $|\eta| < 2.2$ and efficient two-hit coverage in the region $|\eta| < 2.5$. The active elements are n-in-n 100 $\mu m \times 150\ \mu m$ pixels.

In turn, the silicon strip tracker spans in the radial region between 20 cm and 116 cm. It is composed of three different subsystems. The Tracker Inner Barrel and Disks (TIB/TID) consist of 4 layers in the barrel and 3 disks at each end, providing up to 4 $r-\phi$ measurements on a trajectory using 320 μm thick silicon micro-strip sensors with their strips parallel to the beam axis in the barrel and radial on the disks. The Tracker Outer Barrel (TOB), characterized by an outer radius of 116 cm, consists of 6 barrel layers of 500 μm thick micro-strip sensors and provides further 6 $r-\phi$ measurements. The TOB extends in z between ± 118 cm. Beyond this z range the Tracker EndCaps (TEC+ and TEC- where the sign indicates the location along the z axis) cover the region $124\ cm < |z| < 282\ cm$ and $22.5\ cm < |r| < 113.5\ cm$. Each TEC is composed of 9 disks, providing up to 9 ϕ measurements per trajectory. In addition, the modules in the first two layers and rings, respectively, of TIB, TID, and TOB as well as rings 1, 2, and 5 of the TECs carry a second micro-strip detector module which is mounted back-to-back with a stereo angle of 100 mrad in order to provide a measurement of the second co-ordinate (z in the barrel and r on the disks). [5].

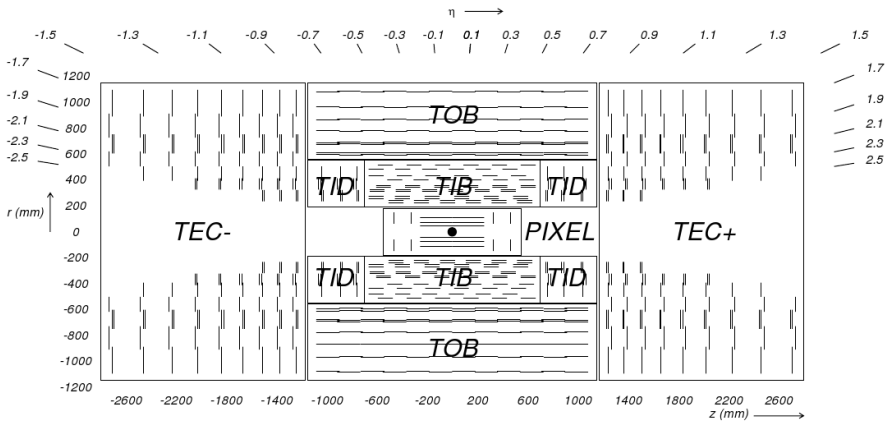

Figure 1.3: Layout of the CMS tracker [5]

1.3 CMS physics results

Since the start of CMS data taking, a number of results have been reached in the physics field covered by the experiment. As figure 1.4 shows, the integrated luminosity has steadily increased between 2010 and 2012, allowing to reach significant amounts of events. Being a general purpose experiment, several particle topics have been studied. Therefore, here only a shortlist of the most relevant outcomes is reported.

The main one is the discovery of a new boson at a mass of 125 GeV [8]. These studies have been performed on proton-proton (p-p) collision data taken with a centre of mass energy $\sqrt{s}$ equal to 7 and 8 TeV. The search has been carried out in five decay channels: $\gamma\gamma$, ZZ, W^-W^+, $\tau^+\tau^-$ and $b\bar{b}$. An excess of events has been observed over the expected background, with a significance of 5.0σ, at a mass around 125 GeV. This was the signal of the production of a new particle. In addition, the measurement of the $\gamma\gamma$ channel is a proof that this particle has a spin different from 1. All these aspects are compatible, within uncertainties, with the expected characteristics of the Standard Model (SM) Higgs boson. Figure 1.5 shows the σ/σ_{SM} quantity, which represents the production cross section times the relevant branching fractions, relative to the SM expectations. The horizontal bars indicate the ± 1 standard deviation uncertainties in the σ/σ_{SM} values for individual modes; they include both statistical and systematic uncertainties.

Another field which has been successfully studied is the B-physics. Among others, a very

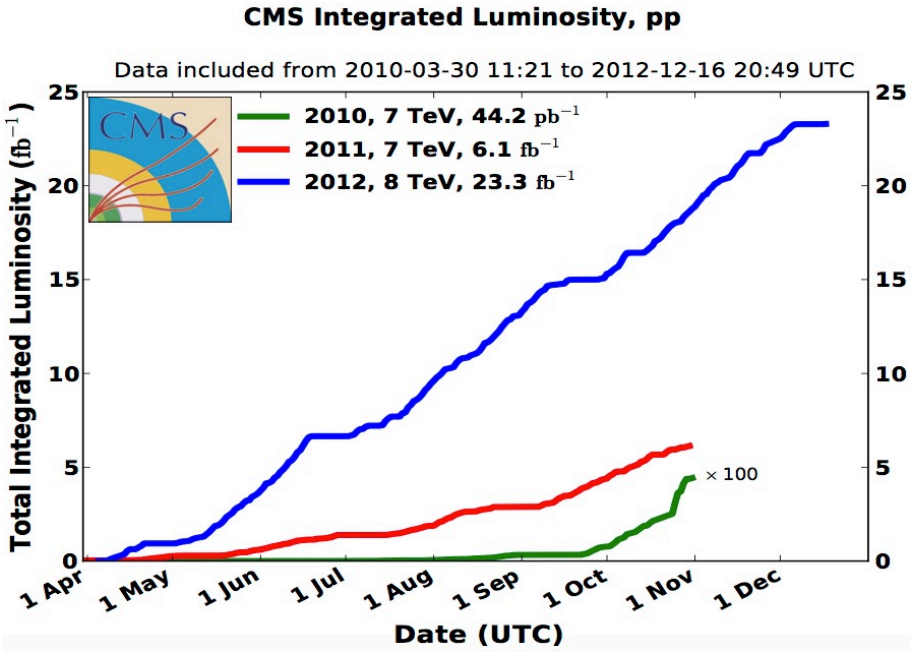

Figure 1.4: Integrated luminosity taken at CMS during Run 1 [7]

important results have been jointly published by CMS and LHCb, experiment specifically designed for this physics field. Based on the Run 1 data taken between 2011 and 2012, the two experiments have observed the very rare decay of the strange B meson $B_S^0 \to \mu^+\mu^-$ [9] with a statistical significance of six standard deviations. The SM predicted this channel with a very low branching fraction. As a results, no evidence of this decay was observed before LHC. In addition, an evidence of the even more rare $B^0 \to \mu^+\mu^-$ decay has been observed with a three sigma significance. Both measurements are compatible with SM predictions, giving some constraints on new physics channels. Figure 1.6 shows the invariant mass $m_{\mu^+\mu^-}$ distribution of the merged CMS and LHCb data. The blue line is the combined fit.
At the moment, no signs of new physics have been found at CMS. Nevertheless, the analysis of Run 1 data has allowed to put some benchmarks on this side. Some theories have been eliminated and others limited. In other cases, the statistics collected in this first step of data taking is not enough to draw conclusions. For these reasons, the Run 2, started in 2015 with an increased center of mass energy and luminosity, is very important, as it will allow to study more in detail the new boson properties, along with beyond SM fields.

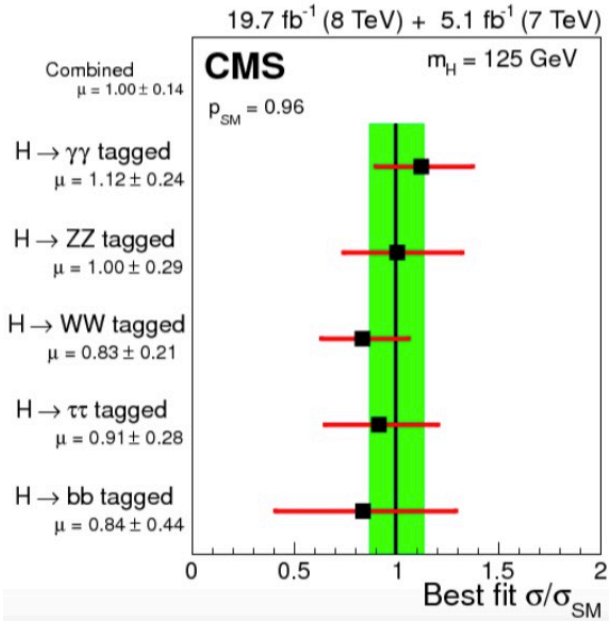

Figure 1.5: Summary of the results for the five channels studied [8]

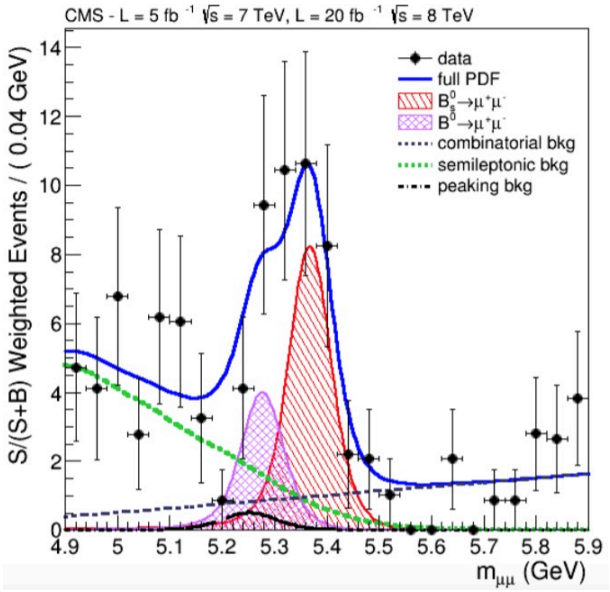

Figure 1.6: Mass distribution for $\mu^+\mu^-$ particles [9]

1.4 HL-LHC upgrade program

As illustrated in section 1.1, after the Long Shutdown 3 (LS3), the HL-LHC will start its operation. The increase in luminosity will be reached thanks to some improvements in the quadrupole magnets and the addition of crab cavities [10]. It is a key improvement regarding studies in some physics channels [11].

- Higgs physics: during the HL-LHC operation more than 100 million Higgs bosons are expected to be produced in CMS. The main channel is gluon-gluon fusion (ggF), followed by vector-boson fusion (VBF), WH, ZH and ttbar H. In this way, a comprehensive study of the Higgs decay modes will be possible. In fact, thanks to the significant increase in statistics, also rare channels can be explored. As an example, the decay $H \to \mu\mu$ will be unambiguously observed and the expected precision in Higgs coupling with muons is 5%. In addition the Higgs boson might also couple to dark matter candidate particles. The branching ratio to invisible decays can be tested with the coupling fits. The resulting expected 95% CL limit on this branching fraction is about 10%. It complements direct searches for invisible decays with similar sensitivity. [11]

- Standard Model tests: one of the most significant topics is the study of Vector Boson Scattering (VBS), like $WW \to WW$. In fact, these processes are important in the study of electroweak symmetry breaking [12]. In addition, searches in the field of the top quark will be performed.

- Beyond the Standard Model: the theories have predicted the existence of new particles at the TeV scale. Nevertheless, up to now, no particles have been observed. Possible causes are that the energy is not sufficient to produce them or that the statistics is too small. The latter aspect will be significantly improved thanks to the HL-LHC machine.

- Flavour physics: the increase in statistics will give the opportunity of improving the precision in the measurement of a number of variables, as an example those of the Cabibbo-Kobayashi-Maskawa (CKM) mechanism. In addition, it should allow significant progresses in the measurement of very rare decay channels, like the $B_{d,s} \to \mu\mu$ decays and about the mixing induced CP violation in B_s^0 decays.

1.5 CMS upgrade programs

The CMS experiment has to get in line with the progressive improvements that LHC is undergoing.

1.5.1 Phase 1 upgrades

Already before HL-LHC, during Phase 1, the different parts of CMS have been experiencing improvements to be able to cope with the increase in luminosity and, therefore, in particle pile-up, i.e. the number of collisions per event. One of the most important intervention is the complete replacement of the silicon pixel detector during the End of Year Technical Stop between 2016 and 2017. In fact, after this pause LHC is expected to deliver an instantaneous luminosity equal to $2 \times 10^{34}\ cm^{-2}s^{-1}$, 2 times the maximum value at which the original pixel detector could operate. In fact, the readout chip would be subject to heavy data losses, which would have a negative impact on the overall CMS performance. The Phase 1 pixel detector design is characterized by [13]:

- Addition of a fourth layer in the barrel and a third disk in the endcap;

- Reduction of the material budget thanks to ultra-lightweight support with CO_2 cooling and disposition of the electronic boards out of the tracking volume;
- Development of a new readout chip with reduced data loss at the Phase 1 luminosity levels.

The barrel additional layer is inserted in order to maintain the previous level of tracking performance even at a higher occupancy. In addition, it can represent a safety margin in case the innermost layer will experience an unexpectedly rapid degradation due to radiation.

1.5.2 Phase 2 upgrades

However, the Phase 2 operation of the CMS experiment will require a further vast upgrade program for the detectors. In fact, they will have to be able to deal with the one order of magnitude increase in radiation (1 Grad in 10 years) and pile-up (between 140 and 200 collisions per event), caused by the significant improvement in luminosity. New choices have to be made in some key general aspects. As an example, the ability to ensure efficient event selection for data acquisition is a key prerequisite to fully benefit from increased luminosity. The precise study of the relatively low-mass Higgs boson discovered in 2012, and the search for new particles occurring in cascade decays will require continued use of low transverse momentum, p_T, trigger thresholds. To achieve this goal, the trigger electronics (i.e. the L1 trigger) must be upgraded [14]. As a consequence, a new approach is required, namely the introduction of tracking information at L1. The upgraded L1 "track trigger" will require a new hardware architecture to incorporate the tracking information. While the addition of track information in the L1 trigger provides significant gains in rate reduction with good efficiency, nevertheless it will be necessary to increase the trigger accept rate in order to maintain the required efficiency for all of the important physics channels.
Another significant improvement in the CMS concept concerns the forward regions of the detector. In fact, a physics acceptance on a wide solid angle is required, in order to improve measurements on processes with small production cross-sections or decays with low branching ratios. The new endcap calorimeter configuration offers the opportunity to extend the muon coverage with a tagging station up to $|\eta| = 3$ or more, with significant acceptance gain for multi-muon final states.
In addition, specific upgrade programs are foreseen for the building blocks of the CMS experiment. The calorimeter endcaps, subject to a significant radiation damage, will be replaced by the High Granularity Calorimeter (HGC). It will be characterized by a fine segmentation in both the transverse and longitudinal directions and will be the first calorimeter specifically optimised for particle flow reconstruction to operate at a colliding beam experiment [15].
Significant changes are foreseen also in the muon endcaps. The current CMS configuration features four stations of Cathode Strip Chambers (CSC)in the region $1.5 < |\eta| < 2.4$. It is the only region of the muon detector that lacks redundant coverage despite the fact that it is a challenging region for muons in terms of backgrounds and momentum resolution. To maintain good L1 muon trigger acceptance in this region it is therefore proposed to enhance these four stations with additional chambers that make use of new detector technologies with higher rate capability, along the lines of what was planned in the original design of CMS.

1.6 Tracker Phase 2 upgrade

The tracking system will experience a large upgrade program, being completely replaced after the LS3. The design itself is renewed in order to take into account the extreme operating conditions of the HL-LHC. To maintain adequate track reconstruction performance at the

much higher pileup levels of the HL-LHC, the granularity of both the outer tracker and the pixel systems will be increased by roughly a factor 4 [16]. The present Outer Tracker was designed to operate without any loss of efficiency up to an integrated luminosity of 500 fb^{-1}, and an average pileup (PU) of less than 50 collisions per bunch crossing. Concerning the pixel detector, despite the complete replacement with the Phase 1 version, the latter restricts the CMS Data Acquisition to a maximum Level-1 (L1) accept rate of about 100 kHz, with an available latency of 4 μs for the trigger decision. It is not compatible with significantly higher rate capability and longer latency at high luminosity. Therefore, a new design has to be conceived. The main requirements that it has to fulfill are:

- **Radiation tolerance**: figure 1.7 shows the map of the expected particle fluence in the Tracker volume corresponding to an integrated luminosity of 3000 fb^{-1}, expressed in terms of 1 MeV neutron equivalent fluence. Figure 1.8, instead, shows detail of the fluence in the pixel volume. In the innermost region it results in a Total Ionizing Dose (TID) of around 1 Grad in 10 years.

- **Increased granularity**: In order to ensure efficient tracking performance at high pileup, the channel occupancy must be maintained near or below the 1% level in all tracker regions, which requires higher channel density. In addition this choice allows an improvement in the resolution of high p_T tracks and in the two-track separation. An average of 140 collisions per bunch crossing is taken as the target number of pileup events to benchmark the performance of the detector.

- **Compliance with the L1 trigger upgrade**: the significant increase in luminosity makes the selection of events at L1 extremely challenging. In addition, the selection algorithms become inefficient at high pile-up. Therefore, in order to maintain or even improve the trigger performance, the maximum L1 rate will be increased to around 1 MHz and the latency to 12.5 μs, adding also the contribution of the tracking detector.

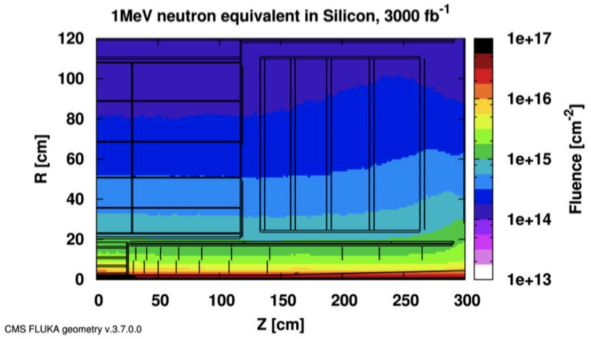

Figure 1.7: Particle fluence in the tracker [14]

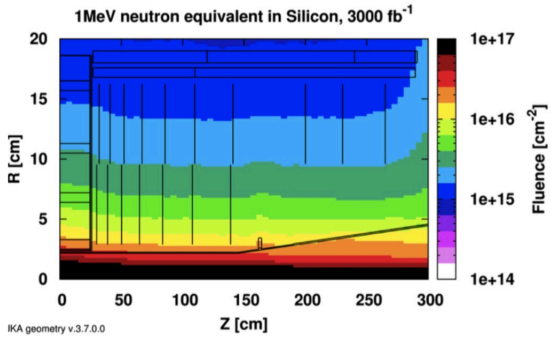

Figure 1.8: Particle fluence in the pixels [14]

Figure 1.9 shows the layout of a quarter of the tracker detector. The outer tracker layout has been subject to intensive studies. The baseline design, presented in the figure, is composed of two different types of modules in the barrel: TBPS, represented in blue, and TB2S in red. The former are composed of two sensors of approximately $5 \times 10\ cm^2$, one segmented in strips, and the other segmented in "macro-pixels" of size $100\ \mu m \times 1.5\ mm$. The latter are composed of two super-imposed strip sensors of approximately $10 \times 10\ cm^2$, mounted with the strips parallel to one another. The same modules are then disposed in vertical direction in the endcap part (TEDD). This choice leads to an efficient use of the silicon sensors while providing good tracking performance while minimizing both cost and material in the tracking volume.

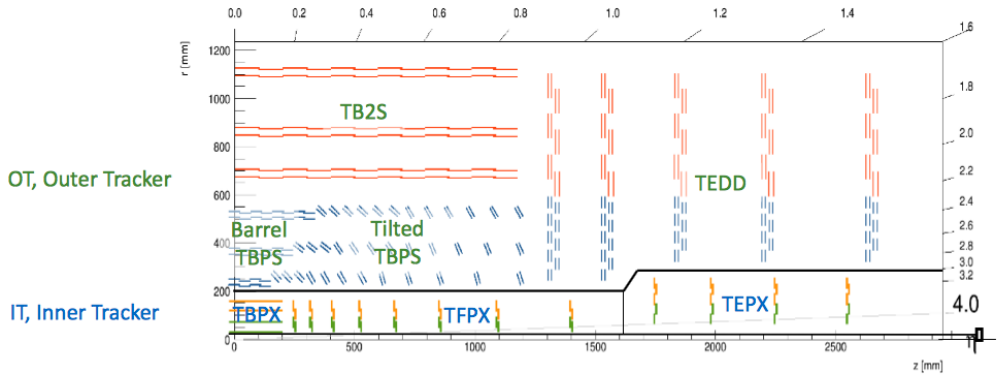

Figure 1.9: Tracker layout for Phase 2

Regarding the pixel detector, it will consist of 4 barrel layers (TBPX) and 12 disks (divided between TFPX and TEPX). The design has been driven by the extreme radiation levels that will be present at HL-LHC. The integrated luminosity of 3000 fb^{-1} results in a particle fluence corresponding to 2×10^{16} n_{eq}/cm^{-2} for the innermost layer located 3 cm away from the interaction region. This value rapidly decreases with the distance. As an example, it is foreseen to be around 3×10^{15} n_{eq}/cm^{-2} at a radius of 11 cm. The geometry of the Phase 2 pixel detector arises from the Phase 1 tracker. In the barrel region 4 layers are foreseen, with an increase in the number of disks from 3 to 12. To cope with the increase in instantaneous luminosity, the pixel area has to be reduced to approximately 2500 μm^2 to keep the occupancy at the percent level [17]. The suggested pixel sizes compatible with this specification are 25×100 μm^2 and 50×50 μm^2. As figure 1.10 shows, it corresponds to a factor six reduction compared to the current barrel pixel size (150×100 μm^2). In this way an improvement in track resolution is obtained, together with a higher robustness with respect to radiation hardness. Monte Carlo simulations are being carried out in order to perform efficiency studies comparing the squared and rectangular solutions [17].

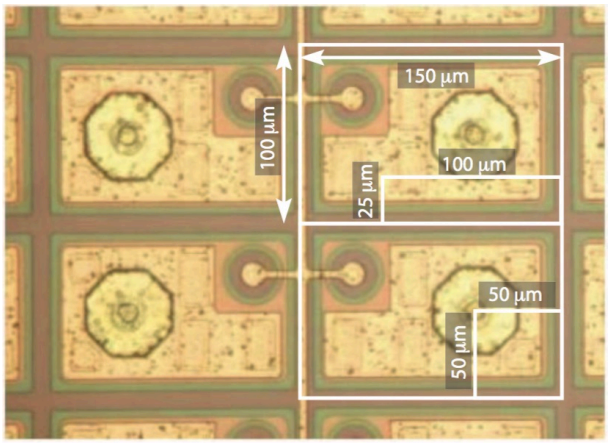

Figure 1.10: Phase 2 pixel dimensions compared to the Phase 1 pixel detector [17] © [2015] IEEE

1.6.1 Sensors

The CMS Phase 2 pixel detector will make use, as for the previous versions of the tracker, of hybrid pixel detectors. As figure 1.11 shows, in this technology the pixelated sensor is bump-

bonded to the pixel chip. In this way the sensors and the chip can be manufactured separately [18].
Silicon sensors have been used since the end of the 1970s as tracking devices in particle physics. The main advantages are the high availability of the material and the fact that an energy of only 3.6 eV is required to produce a electron-hole pair. As a result, also a good energy resolution can be obtained. They are based on the concept of the reverse-biased p-n junction. In this way the depletion region is widened as much as possible. This choice is driven by the fact that in such a region no free carriers are present. When a particle crosses the sensor, it releases some energy in this area producing a certain number of electron-hole pairs. The drift of these charges driven by the external electric field induces the signal on the sensor electrodes. Through the bump bonding the signal is then sent into the electronic readout chain.

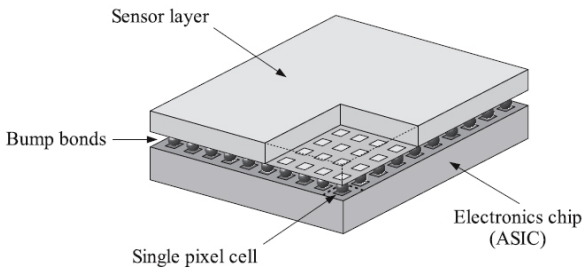

Figure 1.11: Hybrid pixel detector diagram [18]

The sensor is characterized by a number of properties, which have a significant influence on its operation [19]:

- **Full depletion voltage**: it is the voltage needed to extend the depletion zone over the whole thickness of the sensor;
- **Leakage or dark current**: it is the current flowing in absence of external effects if a reverse bias is applied. Until the full depletion is reached the leakage current increases with the square root of the bias. Then a plateau region is entered until the so-called breakdown point is reached at very high voltages, resulting in a sharp increase of the current. As a consequence, the sensor should be operated below the breakdown voltage;
- **Pixel capacitance**: it has a strong influence on the performance of the readout chip. It is composed of several parameters. The first is the capacitance to the backplane, determined by the following expression:

$$C_{back} = \varepsilon_0 \varepsilon_{Si} \frac{A}{d} \tag{1.6.1}$$

  in which A is the pixel area and d the sensor thickness.
  The main contribution is nonetheless given by the interpixel capacitance, i.e. the capacitance to the neighbor pixels. It is proportional to the perimeter. It is also responsible of cross talk between pixels, since a charge deposited on one pixel can induce a signal to the neighboring ones via capacitive coupling. In addition, also the capacitance to the ground plane of the readout chip has to be taken into account.

As explained before, in experiments like CMS, the pixel sensors will be exposed to critical radiation levels, especially in the innermost layers. Radiation induces progressive damage effects in silicon. They can be usually divided into two categories: bulk and surface defects. The former are provoked by the displacement of crystal atoms, while the latter include all

effects near the interface regions. Surface effects have an influence on the charge trapped in the isolation oxides. Bulk damage is caused by the interaction of particle with the nuclei of lattice atoms. Since different kinds of particle at different energies provoke a variety of bulk damages, a physical quantity is needed to be able to compare them. Therefore, radiation damage is scaled using the nonionizing energy loss (NIEL). It summarizes all energy deposited in the crystal which has not been used for the fully reversible process of ionization. The reference particles are 1 MeV neutrons. The main bulk effects are [19]:

- **Increase of the leakage current**: radiation causes a quite sharp increase of the dark current. In numbers it can be expressed as follows [20]:

$$I_{rad} = I_0 + \alpha \Phi A d \tag{1.6.2}$$

  in which I_0 is the dark current before irradiation, α is a damage constant expressed in A/cm, Φ is the radiation fluence, usually given in n_{eq}/cm^2, while $A \cdot d$ expresses the sensor volume. This relationship is valid for the room temperature configuration. Nevertheless, it should be taken into account that leakage current has a strong dependence on temperature:

$$I_{rad}(T) \propto T^2 e^{-\frac{E}{2k_B T}} \tag{1.6.3}$$

  in which $E = 1.2\ eV$ is the activation energy for radiation damaged samples. As a consequence, the pixel detector will be operated at cold temperature, likely around -20 ° C, since under equal radiation levels it results in a leakage current reduced by more than an order of magnitude compared to the room temperature operation.

- Change of the charge in the depleted region, called **type inversion or space charge sign inversion**. It is characterized by a change in the effective doping concentration with radiation. An example of the changes is illustrated in figure 1.12 for a 300 μm thick sensor [19]. The starting n-doped material progressively loses effective doping up to a point around a fluence equal $10^{12}\ n_{eq}/cm^2$ in which it vanishes completely. This is caused by acceptor-like defects which, with further increase of radiation levels, become dominant. As a result, the depletion behavior is like a p-doped material. It results also in a variation of the depletion voltage, since they are bounded by the following relationship:

$$|N_{eff}| = \frac{2\varepsilon_0 \varepsilon_{Si} V_{depl}}{e d^2} \tag{1.6.4}$$

  in which d is the thickness of the sensor. It should be noted that initially p-doped material do not suffer type inversion, since the acceptor defects concentration adds itself to the original doping level.

- **Charge trapping**: most of the radiation-induced defects are unoccupied in the depletion region due to the lack of free carriers. Therefore, they can hold parts of the signal charge carriers for a time significantly larger than the collection time, reducing the signal height. Al low fluences it is a less important issue compared to the previous ones, but it becomes dominant around $10^{15}\ n_{eq}/cm^2$, progressively limiting the operation of silicon sensors in high radiation environment.

Keeping in mind all these aspects, for the Phase 2 upgrade two development programs, respectively on thin n-in-p planar and 3D silicon sensor technologies, have been developed.

Planar silicon sensors

Planar sensors have been developed since the late 1970s. At the beginning, only the p^+-in-n^+ approach was available, i.e. the placement of p^+ implants over a n^+ substrate. This technology

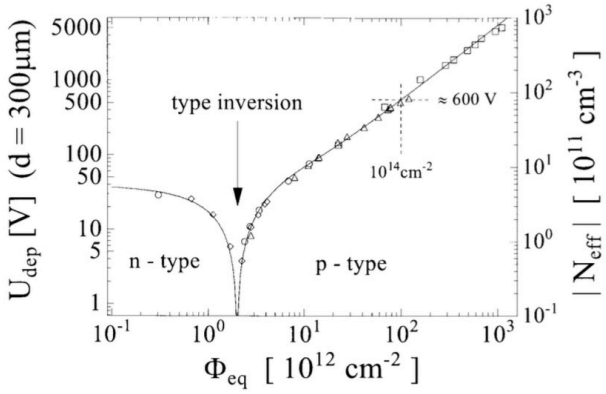

Figure 1.12: Variation of full depletion voltage and effective doping after irradiation [19]

was in fact quite easy to produce. The depletion zone is created by applying a bias voltage on the n-doped backside contact. Nevertheless, it was not able to sustain particle fluences of about $2 \times 10^{14} n_{eq}/cm^2$ [19]. Afterwards, a n^+-in-n approach has been adopted. In this case n^+ implants over a n substrate are implemented, with the junction created with the p^+ backside contact. The advantage of this sensor architecture is that it can be operated in an underdepleted configuration after irradiation. The reduction of the depletion region allows to limit the value of the bias voltage in fact, if it overcomes 1000 V, some issues like large dark currents and cable isolation start to occur. On the other hand, a double-sided processing is mandatory for the junction isolation, resulting in a significant increase in cost and the device is more prone to damage on the backside which destroys the junction [19]. Finally it is also possible to build sensors with the n^+ electrodes in a p substrate approach, illustrated in figure 1.13. n^+-in-p is the most promising choice for planar sensor in a high radiation environment for a number of reasons [21]:

- It does not show type inversion and can be operated partially depleted, in fact the p–n junction is formed on the readout side. As a consequence, also after irradiation the maximum depletion voltage can be probably kept below 1 kV, a voltage which is quite difficult to manage in the detector;
- It needs only single sided lithographic processing and therefore is potentially inexpensive compared with a typical n-in-n type sensor;
- It collects electrons which are, due to their higher mobility, less prone to trapping.

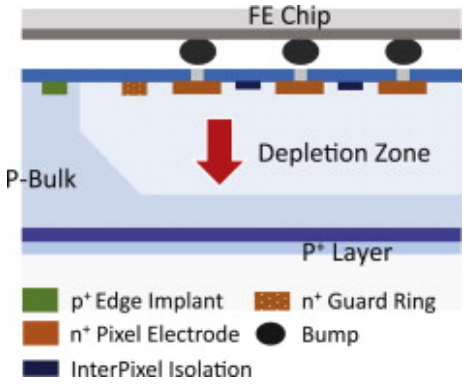

Figure 1.13: Picture of an n-in-p planar pixel sensor [21]

In addition, also the active thickness of the sensor will be reduced, with several advantages [17]:

- High electric fields can be reached with lower voltages, resulting in an adequate charge collection;
- A reduction of the volume and of the voltage leads to smaller bulk currents. Therefore the noise and the power consumption decrease, reducing the impact on the cooling system;
- A smaller active thickness reduces the effects tied to particle trapping;
- High electric fields can lead to charge multiplication caused by impact ionization, increasing then the signal.

On the other hand, a thinner sensor produces a smaller signal for non-irradiated sensors. Intensive radiation studies have shown that the best compromise is represented by a thin planar epitaxial sensor with a 100 μm thickness.

3D silicon sensors

The second possibility is to use 3D sensors [22]. Figure 1.14 shows an example of these devices, which have been developed in the 1990s to overcome the issues of planar sensors related to high particle fluences. These studies found the evidence that a reduced proximity of the electrodes compared to a standard one (for example 300 μm) not only reduces the depletion voltage but also lowers the trapping probability of generated carriers after radiation induced defects are formed, resulting in a reduced degradation of the signal efficiency. In 3D sensors the electrode distance and the substrate thickness (Δ) can be decoupled: the depletion region grows laterally between the electrodes, whose distance is much smaller than the substrate thickness, so that the full depletion voltage can be dramatically reduced with respect to planar sensors. As a result, also the charge collection time is reduced by almost an order of magnitude. The amount of charge generated by a Minimum Ionizing Particle (MIP) is the same for both sensor types if they have the same substrate thickness. Considering operation after irradiation, the small distance between the electrodes allows a full lateral depletion with voltage lower than 200 V, leading to a significant reduction of power dissipation compared to irradiated planar sensors. [23]. The major disadvantages are instead the increase of pixel capacitance, caused by the small distance between the electrodes, and the higher production costs. As a consequence, this technology appears more suitable for the parts of the pixel detector subject to the higher radiation levels.

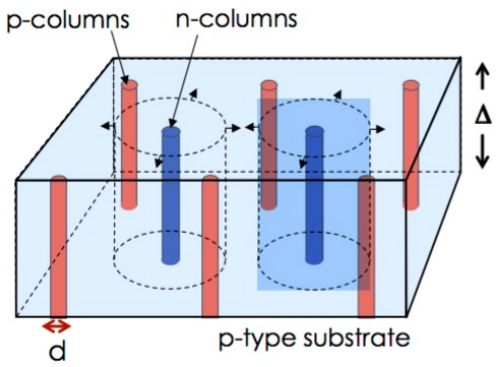

Figure 1.14: 3D sensor [22]

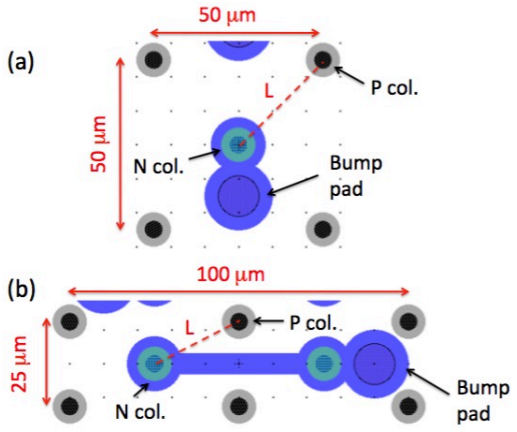

Figure 1.15: Layouts of 25×100 and 50×50 μm^2 3D pixels [22]

Figure 1.15 shows possible implementations of 3D pixel sensors for the Phase 2 upgrade for both desired sizes: $50 \times 50\ \mu m^2$ and $25 \times 100\ \mu m^2$. The former features only one n^+ columns, while the latter two. Therefore, the $25 \times 100\ \mu m^2$ sensor is characterized by smaller inter-electrode spacings (L) making it more radiation tolerant [22]. At the same time, the presence of two readout columns leads to a pixel capacitance around 100 fF, doubled with respect to the $50 \times 50\ \mu m^2$ design. In addition, it makes the placing of the bump bonding quite critical, since it is near to the n^+ and p^+ columns.

1.6.2 Readout Chip

The ReadOut Chip (ROC) for the Phase 2 silicon pixel detector requires a completely new design in order to be able to withstand the extreme conditions present at HL-LHC. In order to develop the new chip, a Research and Development (R & D) collaboration, called RD53, has been developed at CERN. Nineteen institutes from nine countries participate in the collaboration, with almost equal contributions from the two experiments. INFN has a strong involvement in RD53 through the CHIPIX65 project, which includes eight local groups (Bari, Bergamo/Pavia, Lecce, Milano, Padova, Perugia, Pisa and Torino). RD53 is a joint effort between CMS and ATLAS. This choice has been driven by the fact that the requirements for the upgrade of the pixel detector are similar, but not identical, between the two experiments. CMS and ATLAS are, as an example, driven by the same physics and have the same need to trigger on high pileup events [24]. In addition, the TID levels expected in the two detectors is similar, requiring a proper characterization of the technology used for the chip production that has been performed within the RD53 collaboration. Nevertheless, some specifications of the two experiments will be different. As an example, the data rate will be different in the two detectors due to the distance of the innermost layer from the beam pipe (expected to be 3 cm in CMS and 4 cm in ATLAS). Also the trigger latency is expected to be different between the two experiments. For this reason, it is expected that two different versions of the readout chip will be implemented, one for CMS and one for ATLAS, using the RD53 collaboration final chip as a baseline for both.
The specifications used in RD53 have been then driven by the operating conditions of the silicon pixel detectors of CMS and ATLAS at HL-LHC. In the few cases in which the requirements of the two experiments are different, the most severe one has been chosen as a reference. A list of the most important specifications is provided in the table 1.1. The single pixel area in the chip has to be adapted to the pixel sensor one. Therefore, a $50 \times 50\ \mu m^2$ layout has been chosen. The sensor leakage current is foreseen to be lower than 10 nA for all the possible sensor choices. The single pixel hit rate will be around an order of magnitude higher than in the present LHC configuration, leading to a value of 3 GHz/cm^2 for the innermost layer of the detector. Nevertheless, the pixel chip has to be designed in order to minimize hit losses due to the dead time caused by signal processing. The requirement is to maintain this parameter below 1%. The expected trigger rate is 1 MHz, with a latency of 12.5 μs. The latter has an impact on the number of on-chip memories needed to store the data before they are readout. Regarding the radiation damage, the expected TID for the 10 years HL-LHC operation is 1 Grad for the innermost layer. Nevertheless, a radiation tolerance of 500 Mrad is sufficient, since a substitution of the inner tracking system at the half of the period is considered affordable. Taking into account these extreme requirements, it has been decided to use a CMOS 65nm technology for the readout chip design. It is in fact expected to be highly radiation tolerant. In addition, thanks to the shrinking of the transistor sizes, it allows to leave enough room for both analog and digital functions in the small pixel area while keeping the power consumption at a low level. Lastly, it is a long term available technology, which is guaranteed to be available for the whole period of time required by the project (around 10 years). The chosen 65nm process has also allowed the development of small prototypes during the first years of RD53, having

Specification	Value
Technology	CMOS 65nm
Pixel cell	$50 \times 50 \mu m^2$
Sensor leakage current	< 10 nA/pixel
Hit rate	3 GHz/cm^2
Dead time loss	$< 1\%$
Trigger rate	1 MHz
Trigger latency	12.5 μs
Radiation dose	500 Mrad

Table 1.1: Specifications of the RD53 chip

the goal of testing the building blocks of the final chip [25].
The PhD activity presented in this work concerns the design of one of the analog front-ends part of the RD53 program. In chapter 2 an overview of the main aspect of CMOS technologies relevant to analog designs is provided. Chapter 3 is instead dedicated to the common features of analog front-end designed for the readout of silicon sensors for high energy physics. In chapter 4 and 5 a detailed description of the designed analog front-end is provided, together with an overview of the measurement performed on the small prototypes submitted to the foundry. Chapter 6 is dedicated to the large scale prototypes designed by the communities, the CHIPIX65 and RD53A demonstrators, in which the front-end described in this work is included.

Chapter 2

Analog design with deep submicron technologies

In this chapter an overview of analog design with deep submicron CMOS technologies is given. At the beginning of the chapter, a description of the main features of MOS transistors, together with their regions of operations, is provided. Subsequently, the short channel effects, typical of the small-sized deep submicron transistor, are illustrated in detail, toghether with mismatch effects, which are a key aspect to be taken into account in circuit design. The last part of the chapter is dedicated to the radiation-induced effects in CMOS devices.

2.1 General aspects on CMOS technologies

The idea of the metal-oxide-silicon field-effects transistor (MOSFETs) was conceived during the 1920s by J. Lilienfeld [26], even before the bipolar transistor. Nonetheless, due to difficulties in fabrication, this technology started to developed only during the 1960s, firstly as n-type transistors only. Shortly afterwards, also the Complementary-MOS (CMOS) process became available [27], allowing the implementation of both types of transistors on the same substrate. The CMOS technology was quickly adopted by a large number of digital designers due to the big advantages that it offered compared to the bipolar transistor: they burn power only during transitions and need only few devices. In addition, it was soon discovered that CMOS devices were easily scalable, as represented by the Moore's law, and were less expensive than other technologies. At the same time, some groups tried to extend the use of these transistors to analog design [28], given the low cost and the possibility of implementing both analog and digital circuits on the same chip. However, at the beginning the usage of CMOS transistors in analog design was not an ideal choice, due to low speed and high noise figures with respect to bipolar devices. The fact that quickly overturned this situation was that the quick scaling down of transistor sizes, faster than bipolar transistor, resulted in a large increase of speed, making them suitable for both analog and digital designs.
Figure 2.1 shows a simplified cross section of a standard CMOS process, useful to describe its main features. The transistors are built on a common substrate which is usually p-doped, since this kind of wafer is cheaper and easier to produce. In addition, its resistivity is higher than in n-doped substrate, considering that holes have a reduced mobility. This feature helps in keeping low noise propagation across complex Integrated Circuits (ICs). PMOS transistors are then built through a local counter-doping in selected areas, named nwells. In addition, also NMOS transistors need an adjustment of their doping profile in the local pwell. Nevertheless, since no electrical isolation is present between the pwell and the global substrate, the latter must be connected to the most negative potential to avoid the risk of forward biasing the junctions between the electrodes and the bulk in NMOS transistors [29].

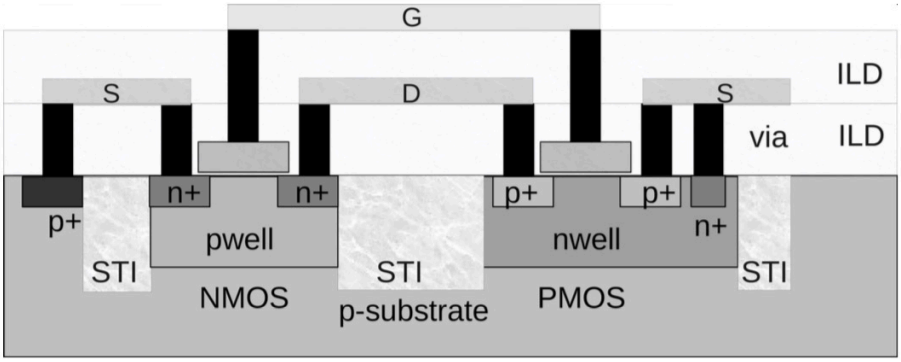

Figure 2.1: Transverse section of a standard CMOS process [29]

In order to guarantee a full electrical isolation between the devices, a structure called Shallow Trench Isolation (STI) is used. It consists of a removal of silicon in areas in which the implementation of transistors is not foreseen, replacing it with oxide. Figure 2.1 also shows that the interconnections between the transistor terminals are implemented using metal lines on different levels, which are joined using vias. The isolation between metals is provided by oxide layers called Inter Level Dielectrics (ILD).
In modern technologies, it is also possible to insert the NMOS transistors into fully isolated wells, realizing the so-called triple well or deep nwell transistor. The structure is depicted in figure 2.2. The first step consists in the implementation of an additional nwell which is deeper than the one used for PMOS transistors. This is why it is named deep nwell. Thereafter, a pwell is created inside the deep nwell in order to create the substrate for the NMOS transistor which is, in this way, fully isolated from the global substrate. To ensure reverse biasing of the junctions, the deep nwell is usually tied to the positive rail [29]. Although this structure requires a larger area with respect to the standard NMOS device, it turns out to be quite effective in very low-noise applications.

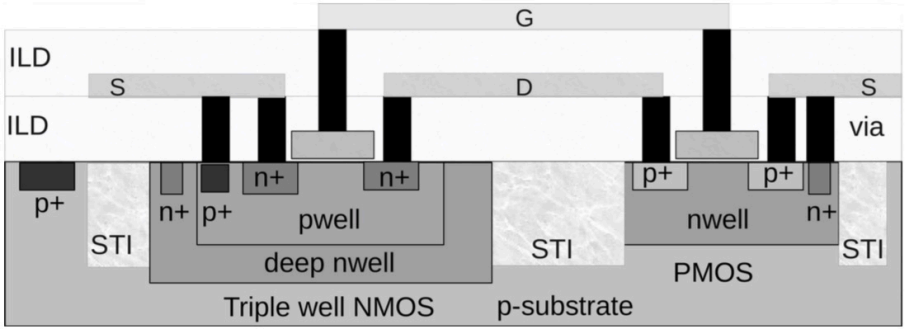

Figure 2.2: Transverse section of a CMOS process with deep nwell NMOS [29]

Lastly, an important role is played by the p^+ substrate contacts, illustrated on the left of both figures 2.1 and 2.2. The reason is the following. The MOS technology is characterized by some intrinsic bipolar transistors, as depicted in figure 2.3. The triggering of these thyristor-like devices leads to a shorting of the power and ground lines, usually resulting in a destruction of the chip, or a system failure that can only be resolved by power-down [30]. This effect is called latchup, and can be simply understood analyzing the equivalent circuit shown in figure 2.4. When one of the two bipolar transistors gets forward biased, for example due to current flowing through the well, it feeds the base of the other transistor. This positive feedback increases the current until the circuit fails or burns out [30]. The way to minimize latchup is to reduce as much as possible the substrate resistivity (i.e. R_{nwell} and R_{psubs}). This goal is reached

using numerous substrate and nwell contacts. In case of a devices carrying large currents, a surrounding line of well/substrate contact, named guard-ring, is required. Usually the rules about substrate contacts are provided by the silicon foundry and should be strictly respected in order to avoid the latchup phenomenon [29].

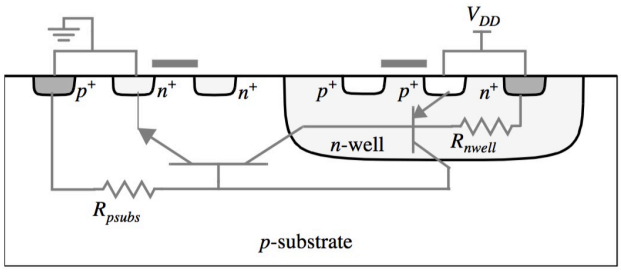

Figure 2.3: Latchup scheme [29]

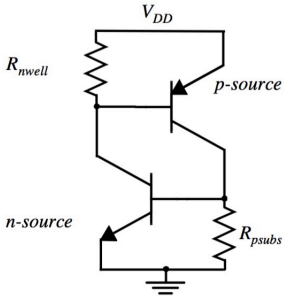

Figure 2.4: Latchup equivalent circuit [29]

Together with the previous aspects, related to the on-silicon fabrication of the devices, a number of electrical properties of the MOS transistors have a direct influence in analog design. An overview of the most important ones is presented in the following paragraphs.

2.2 Threshold voltage

As it will described in section 2.3, the channel in a MOS transistor is formed only in case the gate-source-voltage is superior to a critical limit, called "threshold voltage". To understand the cause of this property, it is useful to model the MOS transistor like a parallel plate capacitor in which the gate and bulk contacts are separated by a layer of insulator (SiO_2). The gate is considered made of n^+ doped silicon and the bulk of p-doped silicon. Then the two terminals can be connected to a battery, as illustrated in figure 2.5. In this model, the threshold voltage

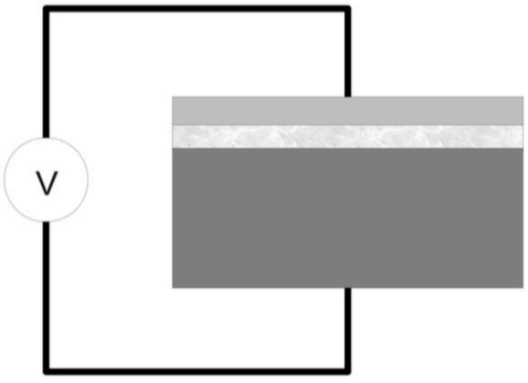

Figure 2.5: MOS capacitor [29]

is represented by the voltage that the battery must apply to attract a significant number of electrons under the gate [29]. Since junctions between different materials are formed, a number of contact potentials is expected. In case of a chain of materials, the difference between the contact potentials of the two extremes of the chain depends only on the properties of the first and last material. As a result, only the value $\phi_G - \phi_B$, in which ϕ_G and ϕ_B are the potentials measured with respect to intrinsic silicon for gate and bulk respectively, has to be considered. Due to the high doping, the gate can be considered degenerate. In other words, the doping level

is so high that the material starts to behave like a metal. As a consequence, it can be found that its surface potential is equal to 0.56 V. For the bulk, supposing a doping profile around $10^{17}\ cm^{-3}$, the contact potential can be evaluated with respect to intrinsic silicon through the following expression:

$$\phi_B - \phi_i = -\frac{kT}{q} ln \frac{N_A}{n_i} = 25.9\ ln \frac{1 \cdot 10^{17}}{1.08 \cdot 10^{10}} = -0.416V \tag{2.2.1}$$

Therefore, $\phi_G - \phi_B$ is a positive quantity and the distribution of charges inside the MOS capacitors in not uniform. An accumulation of negative charges tends to appear at the interface between the bulk and the oxide. In turn, positive charges tend to collect between the gate and the oxide. In addition, the fabrication of the gate oxide may lead to a residual charge trapping in the oxide (Q_{ox}). Since the gate capacitance is given by

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} \tag{2.2.2}$$

in which t_{ox} is the oxide thickness and ε_{ox} is the dielectric constant of SiO_2, the additional voltage contribution due to this effect is given by the ratio between Q_{ox} and C_{ox}. It is usually quite small, around 2 mV. As a consequence, the total voltage that should be supplied by the battery in order to level out the charge distribution is:

$$V_{FB} = \phi_B - \phi_G - \frac{Q_{ox}}{C_{ox}} \tag{2.2.3}$$

V_{FB} is called "flat-band voltage". The name rises from the energy bands description. In fact, the non-uniform distribution of charges inside a structure is represented as a bending of the valence and conduction band. Then, the voltage generated by the battery avoids the bending of the bands inside a structure and guarantees a uniform distribution of charges over the whole sample.
Starting now with the battery charged at the V_{FB} value, it is possible to increase the gate potential in order to understand the MOS behavior. In this way the holes at the bulk-oxide interface are repelled, giving rise to a depletion region containing negatively ionized acceptor atoms. With a further increase of the gate voltage, the depletion region widens and free electrons start to concentrate at the interface. As a result, a part of the bulk contains now charges with the opposite sign with respect to the p doping. Therefore, the surface is "inverted". This process is maintained only as long as the external voltage is provided. In addition, it should be kept in mind that in MOS structures two conditions must be verified. The first one is the energy conservation, given by:

$$V_{GB} = V_{FB} + \psi_{ox} + \psi_{sub} \tag{2.2.4}$$

in which ψ_{ox} is the voltage across the gate oxide and ψ_{sub} is the voltage difference between the bulk-oxide interface and the neutral silicon bulk. In fact, the battery should generate a voltage corresponding to the drops inside the device. The second one is the conservation of charge and the neutrality of the capacitor, stated by:

$$Q_G + Q_{ox} + Q_{ch} = 0 \tag{2.2.5}$$

Since V_{FB} and Q_{ox} are constant quantities, a change in gate voltage leads to the following relationships:

$$\Delta V_{GB} = \Delta\psi_{ox} + \Delta\psi_{sub} \tag{2.2.6}$$

$$\Delta Q_G + \Delta Q_{ch} = 0 \tag{2.2.7}$$

Defining with n_s the electrons concentration inside the channel, it is possible to relate the electrons concentration at the surface of the Si, n_s, and the one in the undisturbed bulk, n_0, to the potential difference between the two regions using the following relationship:

$$V_1 - V_2 = \phi_T \; ln\frac{n_1}{n_2} \tag{2.2.8}$$

which states the concentration of carriers in different parts of the semiconductor is proportional to the voltage difference between the same points. ϕ_T is called "thermal voltage" and is in turn given by:

$$\phi_T = \frac{k_B T}{q_e} \tag{2.2.9}$$

It is equal to 25.9 mV at room temperature. Applying now equation 2.2.8 to this context, it is possible to write:

$$\psi_{sub} = \phi_T \; ln\frac{n_s}{n_0} \tag{2.2.10}$$

Therefore n_s is equal to:

$$n_s = n_0 \; e^{\frac{\psi_{sub}}{\phi_T}} = \frac{n_i^2}{p_0} e^{\frac{\psi_{sub}}{\phi_T}} \tag{2.2.11}$$

in which the law of mass action ($n_0 p_0 = n_i^2$) has been used. Some more considerations can be drawn by using the Fermi potential of the bulk (a full dissertation can be found in [26]) which is given by:

$$\phi_F = \phi_T \; ln\frac{p_0}{n_i} \tag{2.2.12}$$

Extracting n_i from the previous relationship it is possible to reshape equation 2.2.11 recalling that p_0 is equal to the acceptor concentration N_A:

$$n_s = N_A \; e^{\frac{\psi_{sub} - 2\phi_F}{\phi_T}} \tag{2.2.13}$$

Therefore, the electrons concentration at the surface depends on the bulk Fermi potential. This relationship leads to some considerations:

- If $\boldsymbol{\psi_{sub} < 2\phi_F}$ the induced electrons concentration is quite modest.
- If $\boldsymbol{\psi_{sub} = 2\phi_F}$ the electrons concentration matches the hole one in the undisturbed bulk. As a result, $2\phi_F$ is the minimum voltage to be applied in order to create an inversion layer and has to be added to the final threshold voltage expression.
- If $\boldsymbol{\psi_{sub} > 2\phi_F}$ the electrons concentration quickly overcomes N_A due to the exponential dependence. Therefore small variations in ψ_{sub} result in quite large variations of n_s.

At this point a last contribution to the threshold voltage has to be considered. In fact, the n^+ layer induced in the p substrate results in an additional n^+p junction, which disappears once the gate voltage voltage responsible of the inversion layer is removed. In order to calculate the charge associated to the junction depletion layer it is possible to assume that since the electrons concentration is larger than the holes one, the depletion region is mainly extended in the p substrate. In addition, a ψ_{sub} exceeding $2\phi_F$ of some tens of millivolts is enough to have a large concentration of electrons. Therefore the substrate potential can be considered constant and equal to $2\phi_F$ at first glance. Considering the extension of the depletion region in case of an asymmetric junction, it is possible to write:

$$x_{dp} = \sqrt{\frac{2\varepsilon_{Si}}{qN_A} 2\phi_F} \tag{2.2.14}$$

Defining W and L as the width and length of the channel, the total charge in the depletion region is:

$$Q_{B,tot} = qN_A WLx_{dp} \tag{2.2.15}$$

Let's consider now the charge per unit area, inserting the expression of x_{dp}:

$$Q_B = qN_A x_{dp} = \sqrt{2\varepsilon_{Si} qN_A 2\phi_F} \tag{2.2.16}$$

Therefore, to maintain the storage of this charge a voltage equal to Q_B/C_{ox} should be applied. Putting everything together, the voltage needed to maintain the inversion layer is given by:

$$V_{TH0} = V_{FB} + 2\phi_F + \frac{1}{C_{ox}}\sqrt{2\varepsilon_{Si} qN_A 2\phi_F} \tag{2.2.17}$$

It is labelled as V_{TH0} because other contributions have to be taken into account in case the source and bulk voltages are different, as illustrated in section 2.3.1.

2.3 Regions of operation

In order to describe analog circuits based on CMOS devices, it is important to know which are their regions of operation. They basically depend on the values of the voltage applied to the terminals.
The configuration in which source, drain and bulk are connected to zero and the gate to a voltage supply can be considered as a starting condition. As explained in the previous paragraph, the gate and the channel can be considered as the two arms of a parallel plate capacitor. As a result, a charge applied to the gate terminal induces a charge equal in magnitude and opposite in sign inside the channel. In fact, in a MOS the application of an external voltage does not create charges, but just moves them from one place to another, leading to the following relationship:

$$\Delta Q_G + \Delta Q_{ch} = 0 \tag{2.3.1}$$

Based on this assumption, it is possible to identify three different regions of operation [29]:

- **Accumulation**: it happens when a negative voltage is applied to the gate. In this case, holes are attracted in the channel region. Therefore, no current can flow between source and drain. The device behaves like a capacitor with the gate and bulk acting as terminals.
- **Depletion**: if the gate voltage rises above zero, holes are driven off the channel. The latter features instead a layer of negatively ionized atoms, which mirror the gate charge. Since they are fixed atoms, no current will be present even applying a voltage difference between source and drain.
- **Inversion**: if the gate voltage continues to increase, at some point free electrons appear in the channel and start to outnumber the fixed atoms. As a result, the concentration of electrons in the channel tends to match the one into the source and drain terminals.

In the following paragraphs, the source potential will be chosen as the reference voltage. Therefore, the gate-source voltage V_{GS} and the drain-source voltage V_{DS} will be used.

2.3.1 Classical MOS characteristics

In classical MOS models, the electrons density in the channel is considered relevant only when $V_{GS} > V_{TH}$. As a result, if the latter condition is in place, the device is considered turned on, otherwise is off. This is a quite huge approximation. In fact, as illustrated in section 2.3.3, a current can be present inside the channel also if $V_{GS} < V_{TH}$. Nevertheless, the classical

approach is useful in order to explore the basic operation of the NMOS transistor. The characteristics of the drain-source current I_{DS} are expressed with respect to the terminal voltages. The difference $V_{GS} - V_{TH}$ is usually referred to as "overdrive voltage". Considering the trend of I_{DS} versus V_{DS} it is possible to identify two different behaviors:

- **Linear or triode region**: the device works in this region if $V_{DS} < (V_{GS} - V_{TH})$. The expression of the current is the following:

$$I_{DS} = \mu_n C_{ox} \frac{W}{L} (V_{GS} - V_{TH}) V_{DS} \tag{2.3.2}$$

  Therefore, as the relationship between I_{DS} and V_{DS} is linear (see figure 2.6 [31]), the transistor behaves like a voltage controlled linear resistor of value:

$$R_{ON} = \frac{V_{DS}}{I_{DS}} = \frac{1}{\mu_n C_{ox} \frac{W}{L} (V_{GS} - V_{TH})} \tag{2.3.3}$$

- **Saturation region**: it is characterized by a $V_{DS} \geq (V_{GS} - V_{TH})$. For the points inside the channel in which this condition is verified, the inversion layer does not form. As a result, this part of the channel does not contribute to the current flow with its own charges. The channel is when said "pinched-off". At the same time, the charges formed in the inverted portion of the channel can cross this region and reach the terminal. In this case the current saturates, as shown in figure 2.6, to the value:

$$I_{DS} = \frac{1}{2} \mu_n C_{ox} \frac{W}{L} (V_{GS} - V_{TH})^2 \tag{2.3.4}$$

  Therefore, in this region the I_{DS} is independent of V_{DS} and the device behaves like a voltage controlled current source. In addition, the voltage $V_{GS} - V_{TH}$ is usually referred to as "saturation voltage".

Figure 2.6 also shows that between the two regions there is a transition interval. As a consequence, in order to use the devices steadily in saturation, it is recommended to increment the V_{DS} above the saturation value by some margin. Depending on the technology, it can be between 0.1 and 0.2 V. A detailed illustration of the topic, comprehensive of a full derivation of the previous relationships can be found in [29][31][32].

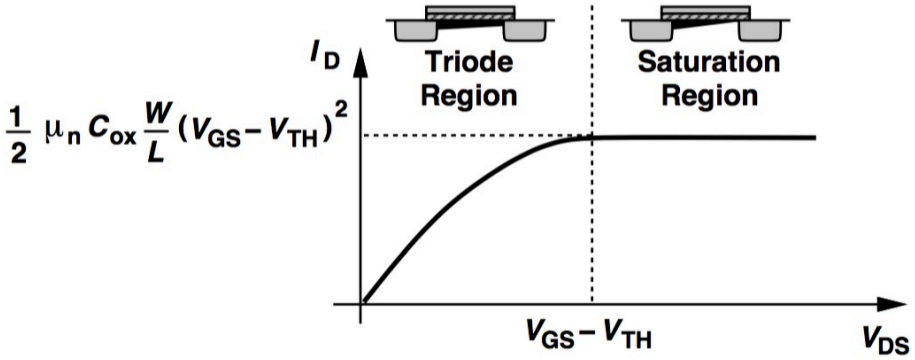

Figure 2.6: I_{DS} versus V_{DS} [31]

Channel length modulation

Equation 2.3.4 states that in the saturation region the I_{DS} is independent on V_{DS}. Nevertheless, this is only an approximation. In fact, as sketched in the previous paragraph, the actual

length of the active channel, L', decreases with the V_{DS}. This effect is called "channel length modulation" [27]. Therefore, it is possible to write [29]:

$$L' = L - \Delta L \tag{2.3.5}$$

in which ΔL is the pinched-off portion in the channel. It can be introduced in equation 2.3.4:

$$I_{DS} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L-\Delta L}(V_{GS} - V_{TH})^2 \tag{2.3.6}$$

which can be modified as follows:

$$I_{DS} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L(1-\frac{\Delta L}{L})}(V_{GS} - V_{TH})^2 = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2\left(1 + \frac{\Delta L}{L}\right) \tag{2.3.7}$$

Since ΔL is smaller than L, in equation 2.3.7 the following approximation has been used:

$$\frac{1}{1-x} \simeq 1 + x \tag{2.3.8}$$

It is then possible to assume a linear relationship between $\frac{\Delta L}{L}$ and V_{DS}:

$$\frac{\Delta L}{L} = \lambda V_{DS} \tag{2.3.9}$$

λ is called "channel length modulation parameter". Finally, the I_{DS} can be expressed as follows:

$$I_{DS} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2(1 + \lambda V_{DS}) \tag{2.3.10}$$

Body effect

In the previous expressions it was assumed that the potentials of the bulk and the source terminals were tied to ground. However it is frequent in circuit applications that for some transistors the source terminal is tied to a voltage $V_S > 0$ with the bulk kept at ground. In this case, the source-substrate junction remains reverse-biased and the device still operates properly [31]. Nevertheless, it can be demonstrated that as the source becomes more positive with respect to the substrate, the threshold voltage increases. This phenomenon is called "body effect" and is described by the following relationship:

$$V_{TH} = V_{TH0} + \gamma(\sqrt{|\,2\phi_F + V_{SB}\,|} - \sqrt{|\,2\phi_F\,|}) \tag{2.3.11}$$

in which V_{SB} indicates the source-bulk potential difference and γ represents the "body effect coefficient" evaluated as:

$$\gamma = \frac{\sqrt{2\varepsilon_{Si} q N_A}}{C_{ox}} \tag{2.3.12}$$

A typical value for γ is 0.5 V $^{1/2}$.

PMOS transistor characteristics

As explained in section 2.1, the PMOS transistor has a complementary behavior with respect to the NMOS device. The bulk is n-doped and the source and drain terminals are p-doped. In order to bring holes in the channel, the gate voltage has to be more negative than the source one by a threshold voltage factor at least. Therefore, the threshold voltage for PMOS transistors is considered negative. However, in order to deal with positive quantities, for PMOS devices the sense of measurement of the voltages in the device is inverted (V_{SG} and V_{DS} are used). In

addition, the absolute value of the threshold voltage is taken. As a consequence, the PMOS characteristics in linear region is the following:

$$I_{SD} = \mu_p C_{ox} \frac{W}{L} \left[(V_{SG} - |V_{THP}|) V_{SD} - \frac{V_{SD}^2}{2} \right] \tag{2.3.13}$$

while in saturation:

$$I_{SD} = \frac{1}{2} \mu_p C_{ox} \frac{W}{L} (V_{SG} - |V_{THP}|)^2 (1 + \lambda V_{SD}) \tag{2.3.14}$$

2.3.2 Small signal parameters

Together with the channel current, other parameters have to be considered in analog design. In fact, the relationships described in the previous paragraph are nonlinear. Nevertheless, a proper linearization of the large-signal equations at a defined operating point is useful for circuit analysis, giving rise to the so called small-signal parameters [33]. In other words, the small signal parameters always relate variations of quantities in the vicinity of a point and not their values at that point.

Gate transconductance

The most important small-signal parameter is the gate transconductance, defined as:

$$g_m = \frac{\partial I_{DS}}{\partial V_{GS}} \tag{2.3.15}$$

It acts as a measure of the modification of the drain current due to variations of the gate-source voltage. Using equation 2.3.4, it can be derived the value of g_m in the saturation region:

$$g_m = \mu_n C_{ox} \frac{W}{L} (V_{GS} - V_{TH}) = \sqrt{2 \mu_n C_{ox} \frac{W}{L} I_{DS}} \tag{2.3.16}$$

Therefore, it has a linear dependence on V_{GS} and is proportional to I_{DS}. It is expressed in Ω^{-1} or Siemens.

Bulk transconductance

As explained in section 2.3.1, the body effect has a significant influence in the threshold voltage, and therefore on the I_{DS}. As a result, another important parameter is the bulk transconductance [27]:

$$g_{mb} = \frac{\partial I_{DS}}{\partial V_{BS}} = \frac{\partial I_{DS}}{\partial V_{TH}} \frac{\partial V_{TH}}{\partial V_{BS}} = -\frac{\partial I_{DS}}{\partial V_{TH}} \frac{\partial V_{TH}}{\partial V_{SB}} \tag{2.3.17}$$

From equation 2.3.11:

$$\frac{\partial V_{TH}}{\partial V_{SB}} = \frac{\gamma}{2} (2\phi_F + V_{SB})^{-1/2} \tag{2.3.18}$$

As a result, the bulk transconductance can be expressed as follows:

$$g_{mb} = \mu_n C_{ox} \frac{W}{L} (V_{GS} - V_{TH}) \frac{\gamma}{2} (2\phi_F + V_{SB})^{-1/2} = g_m \frac{\gamma}{2} (2\phi_F + V_{SB})^{-1/2} = \eta g_m \tag{2.3.19}$$

in which the parameter η is usually between 0.2 and 0.3. This equation also shows that g_{mb} is proportional to γ and that the incremental body effect becomes less pronounced as V_{SB} increases.

Output impedance

As illustrated in section 2.3.1, the channel length modulation makes the drain-source current dependent on the V_{DS}. This effect can be modeled by a voltage dependent current source. In other words, a current source having a value linearly dependent on the voltage across it behaves like a resistor tied between D and S [27]. The value of the output resistance can be found starting from equation 2.3.10 :

$$r_0 = \frac{\partial V_{DS}}{\partial I_{DS}} = \frac{1}{\frac{\partial I_{DS}}{\partial V_{DS}}} = \frac{1}{\frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS}-V_{TH})^2\lambda} \simeq \frac{1}{\lambda I_{DS}} \tag{2.3.20}$$

The reciprocal is the output conductance:

$$g_{ds} = \frac{1}{r_0} = \lambda\mu_n C_{ox}\frac{W}{L}(V_{GS}-V_{TH}) \tag{2.3.21}$$

The quantity $g_m r_0$ is defined as "intrinsic gain". In fact, with no external loads, it represents the maximum voltage gain provided by a single transistor [31].

2.3.3 Subthreshold conduction

The approximation used in section 2.3.1, stating that no current flows in the channel if $V_{GS} < V_{TH}$ is quite ineffective, especially in modern deep submicron technologies. In fact, accurate models show that below threshold the drain-source current has an exponential dependence on the gate voltage. It is referred to as "subthreshold or weak inversion" region. At the same time the case $V_{GS} > V_{TH}$ is called "strong inversion" region [28]. Since the weak inversion conduction is relevant to the analog design described in this work a detailed derivation is presented here. It can be started by finding the mobile charge in the channel solving the Poisson equation [29]:

$$\frac{d^2\psi_s}{d^2y} = -\frac{\rho}{\varepsilon_{Si}} \tag{2.3.22}$$

in which ψ_s is the surface potential, ρ the charge density and ε_{Si} the silicon dielectric constant:

$$\rho = -n_0 e^{\frac{\psi_s(y)}{\phi_T}} + p_0 e^{-\frac{\psi_s(y)}{\phi_T}} - N_A \tag{2.3.23}$$

in which n_0 and p_0 are the electrons and holes concentrations in the bulk. Beyond depletion, the term relevant to the holes can be neglected. Therefore, solving equation 2.3.22 with this assumption, the following result is obtained:

$$Q_{ch} = -\sqrt{2\varepsilon_{Si}qN_A}\sqrt{\psi_s + \phi_T e^{\frac{(\psi_s - 2\phi_F)}{\phi_T}}} \tag{2.3.24}$$

in which ϕ_F is the Fermi potential. Q_{ch} represents the charge density per unit area in a specific point of the channel. In addition, it can be expressed as the sum of the mobile charge Q_M due to electrons and the fixed charge Q_d due to ionized acceptor atoms.

$$Q_{ch} = Q_M + Q_d \tag{2.3.25}$$

Taking into account that the charge due to acceptor ions is negative, the charge density due to the depletion layer is:

$$Q_d = -\sqrt{2\varepsilon_{Si}qN_A\psi_s} \tag{2.3.26}$$

As a result, it is possible to extract the value of Q_M using the previous relationships, obtaining:

$$Q_M = Q_{ch} - Q_d = -\sqrt{2\varepsilon_{Si}qN_A}\left(\sqrt{\psi_s + \phi_T e^{\frac{(\psi_s - 2\phi_F)}{\phi_T}}} - \sqrt{\phi_s}\right) \tag{2.3.27}$$

If the term $\phi_T e^{\frac{(\psi_s - 2\phi_F)}{\phi_T}}$ is small, it is possible to perform a first order Taylor approximation of the first term in the parenthesis, based on the rule:

$$\sqrt{x+y} \simeq \sqrt{x} + \frac{y}{2\sqrt{x}} \tag{2.3.28}$$

As a consequence, equation 2.3.27 can be rewritten as follows:

$$Q_M = -\frac{\sqrt{2\varepsilon_{Si} q N_A}}{2\sqrt{\psi_s}} \phi_T e^{\frac{(\psi_s - 2\phi_F)}{\phi_T}} \tag{2.3.29}$$

Furthermore, the dependency on the channel bulk voltage V_{CB} can be added in equation 2.3.29 replacing $2\phi_F$ with $2\phi_F + V_{CB}$:

$$Q_M = -\frac{\sqrt{2\varepsilon_{Si} q N_A}}{2\sqrt{\psi_s}} \phi_T e^{\frac{(\psi_s - 2\phi_F - V_{CB})}{\phi_T}} = -\frac{\sqrt{2\varepsilon_{Si} q N_A}}{2\sqrt{\psi_s}} \phi_T e^{\frac{(\psi_s - 2\phi_F)}{\phi_T}} e^{-\frac{V_{CB}}{\phi_T}} \tag{2.3.30}$$

Diffusion is the main mechanism behind the current flowing in weak inversion. In fact, the electrons that cross the potential barrier at the source potential are then located in a space with very few other electrons around. As a result, they start diffusing towards the drain. Assuming a linear gradient from the source to the drain, the diffusion current can be written as:

$$I_{DS} = -\mu_n \phi_T W \left(\frac{Q_S - Q_D}{L} \right) = -\mu_n \phi_T \frac{W}{L} Q_S \left(1 - \frac{Q_D}{Q_S} \right) \tag{2.3.31}$$

in which Q_D and Q_S are the charge per unit area for source and drain respectively. Using 2.3.30 it is possible to write:

$$Q_S = -\frac{\sqrt{2\varepsilon_{Si} q N_A}}{2\sqrt{\psi_s}} \phi_T e^{\frac{(\psi_s - 2\phi_F)}{\phi_T}} e^{-\frac{V_{SB}}{\phi_T}} \tag{2.3.32}$$

$$Q_D = -\frac{\sqrt{2\varepsilon_{Si} q N_A}}{2\sqrt{\psi_s}} \phi_T e^{\frac{(\psi_s - 2\phi_F)}{\phi_T}} e^{-\frac{V_{DB}}{\phi_T}} \tag{2.3.33}$$

As a result:

$$\frac{Q_D}{Q_S} = e^{-\frac{V_{DS}}{\phi_T}} \tag{2.3.34}$$

Analyzing in detail the Q_S expression, some assumptions can be made. The quantity before the exponential has a weak dependence of ψ_s, as it appears under the square root. It is then possible to approximate ψ_s with $2\phi_F + V_{SB}$. In fact, the channel is considered inverted if $\psi_s = 2\phi_F$ or if $\psi_s = 2\phi_F + V_{SB}$ in case source and bulk are not at the same potential. As a result, the depletion layer charge can be expressed as follows:

$$Q_d = \sqrt{2\varepsilon_{Si} q N_A (2\phi_F + V_{SB})} \tag{2.3.35}$$

The depletion layer capacitance is:

$$C_d = \frac{dQ_d}{dV_{SB}} = \frac{\sqrt{2\varepsilon_{Si} q N_A}}{2\sqrt{2\phi_F + V_{SB}}} \tag{2.3.36}$$

It is now possible to introduce a new quantity, called "slope factor", which is defined as follows:

$$n = 1 + \frac{C_d}{C_{ox}} \tag{2.3.37}$$

Taking now the factor before the exponential in equation 2.3.32:

$$Q_{s0} = -\frac{\sqrt{2\varepsilon_{Si} q N_A}}{2\sqrt{\psi_s}} \phi_T \tag{2.3.38}$$

As already done for equation 2.3.35, ψ_s can be approximated with $\psi_s = 2\phi_F + V_{SB}$:

$$Q_{s0} = -\frac{\sqrt{2\varepsilon_{Si} q N_A}}{2\sqrt{2\phi_F + V_{SB}}}\phi_T \tag{2.3.39}$$

Multiplying and dividing now 2.3.39 by C_{ox} and using 2.3.36 and 2.3.37 it is possible to write:

$$Q_{s0} = -C_{ox}\phi_T(n-1) \tag{2.3.40}$$

Inserting 2.3.40 and 2.3.33 in equation 2.3.31 we have:

$$I_{DS} = \mu_n C_{ox}(n-1)\frac{W}{L}\phi_T^2 e^{\frac{\psi_s - 2\phi_F - V_{SB}}{\phi_T}}\left(1 - e^{-\frac{V_{DS}}{\phi_T}}\right) \tag{2.3.41}$$

Since the exponential is a strong function of ψ_s the approximation $\psi_s = 2\phi_F + V_{SB}$ has not been performed. In weak inversion, the charge in the channel is not sufficient to shield the bulk charges and from the gate one sees two capacitors in series, the gate oxide capacitance C_ox and the depletion layer capacitance C_d. Therefore, the voltage which is capable of attracting charges is given by the capacitor divider rule. The following approximation can be made:

$$\psi_s - 2\phi_F - V_{SB} \simeq (V_{GS} - V_{TH}(V_{SB}))\frac{C_{ox}}{C_d + C_{ox}} = \frac{(V_{GS} - V_{TH}(V_{SB}))}{n} \tag{2.3.42}$$

As a consequence, the final expression of the drain-source courrent in weak inversion can be written as follows:

$$I_{DS} = \mu_n C_{ox}(n-1)\frac{W}{L}\phi_T^2 e^{\frac{V_{GS} - V_{TH}}{n\phi_T}}\left(1 - e^{-\frac{V_{DS}}{\phi_T}}\right) \tag{2.3.43}$$

Equation 2.3.43 shows how different is the expression of the current in weak inversion compared to the strong inversion case.

inversion coefficient and g_m/I_D methodology

As shown in previous section, the derivation of equations that account for the MOS behavior across all its possible range of operation only from first principles is very difficult. As a consequence, the development of models able to describe the complexity of real transistors is crucial. An important example is the EKV model [33]. It has been developed keeping in mind the basic purpose that a model should provide: a good understanding of the various properties of the device to facilitate the synthesis of optimum circuit architectures and an easy adaptation to numerical simulations on a computer, embedded in a circuit simulator. The EKV is referenced to the local substrate and not to the source, and it assumes a complete symmetry between source and drain. In absence of channel length modulation, the EKV model describes the weak inversion behavior as follows:

$$I_{DS} = 2n\mu C_{ox}\frac{W}{L}\Phi_T^2 e^{\frac{V_{GS} - V_{TH}}{n\Phi_T}} \tag{2.3.44}$$

The strong inversion current is instead given by:

$$I_{DS} = \frac{1}{2}\mu C_{ox}\frac{W}{nL}(V_{GS} - V_{TH})^2 \tag{2.3.45}$$

Since these expressions are part of the same model, it is possible to try to derive quantities in order to predict the region of operation of a device. This goal can be achieved by finding a boundary current between weak and strong inversion. It is defined as the current for which

the value of g_m found using equation 2.3.44 is equal to the one derived from equation 2.3.45 [29]. Knowing that in weak inversion the transconductance is given by:

$$g_m = \frac{\partial I_{DS}}{\partial V_{GS}} = \frac{1}{n\Phi_T} 2n\mu C_{ox} \frac{W}{L} \Phi_T^2 e^{\frac{V_{GS}-V_{TH}}{n\Phi_T}} = \frac{I_{DS}}{n\Phi_T} \tag{2.3.46}$$

it is then possible to write:

$$\frac{I_{DS,boundary}}{n\Phi_T} = \sqrt{2\mu C_{ox} \frac{W}{nL} I_{DS,boundary}} \tag{2.3.47}$$

The boundary current is therefore given by:

$$I_{DS,boundary} = 2n\mu C_{ox} \frac{W}{L} \phi_T^2 \tag{2.3.48}$$

Starting from this relationship it is possible to define the Inversion Coefficient (I_C) as the ratio between the bias current at which the device operates and its boundary current:

$$I_C = \frac{I_{DS}}{2n\mu C_{ox} \frac{W}{L} \phi_T^2} \tag{2.3.49}$$

Based on the value of the inversion coefficient, it is possible to know the region of the operation of the considered device:

- A $I_C < 0.1$ corresponds to the weak inversion regime;
- A $I_C > 10$ corresponds to the strong inversion regime;
- The interval $0.1 < I_C < 10$ marks a transition region between the two configurations, which is called moderate inversion.

Is is therefore possible to express all the main quantities as a function of the inversion coefficient. As an example, the transconductance is given by:

$$g_m = \frac{I_{DS}}{n\phi_T} \frac{1}{\sqrt{I_C + 0.5\sqrt{I_C} + 1}} \tag{2.3.50}$$

This concept is also useful to define the quantity γ, called "inversion factor":

$$\gamma = \frac{1}{2} + \frac{1}{6} \frac{I_C}{I_C + 1} \tag{2.3.51}$$

It is mainly used to quantify the effect of the degree of channel inversion on the thermal noise generated in the device channel [29], and will be used in the next chapters in which noise is discussed. In weak inversion, in which I_C tends to zero, $\gamma \simeq 0.5$, while in strong inversion $\gamma \simeq 2/3$.
Another method frequently used for a unique treatment of all regions of operation is a g_m/I_D based methodology, first presented in 1996 in [34]. This model considers the relationship between the g_m/I_D ratio and the normalized drain current $I_D/(W/L)$ as a fundamental design tool. The g_m/I_D has been chosen because it is strongly related to the performances of analog circuits. In addition, it gives an indication of the device operating region and provides a tool for calculating the transistor dimensions. This ratio is a measurement of the efficiency of translating current into transconductance. In addition it is equal to the derivative of the logarithm of I_D with respect to the gate voltage:

$$\frac{g_m}{I_D} = \frac{1}{I_D} \frac{\partial I_D}{\partial V_G} = \frac{\partial ln\left[\frac{I_D}{\frac{W}{L}}\right]}{\partial V_G} \tag{2.3.52}$$

The derivative is maximum in the weak inversion region where the drain current dependence versus the gate voltage is exponential while it is quadratic in strong inversion. The g_m/I_D ratio decreases as the operating point moves toward strong inversion when I_D or V_G are increased. In addition, it is interesting to analyze the behavior of this ratio with respect to the inversion coefficient. As equation 2.3.52 shows, the g_m/I_D is independent on the transistor aspect ratio. Therefore the relationship between the g_m/I_D and I_C is a unique characteristic for all transistors of the same type (NMOS or PMOS) on the same production batch. In addition, by knowing two parameters between g_m/I_D, g_m and I_D the W/L of a transistor can be unambigously determined. As shown in figure 2.7, in weak inversion the g_m/I_D as a function of I_C remains almost constant, confirming the direct proportionality between g_m and I_D in this region.

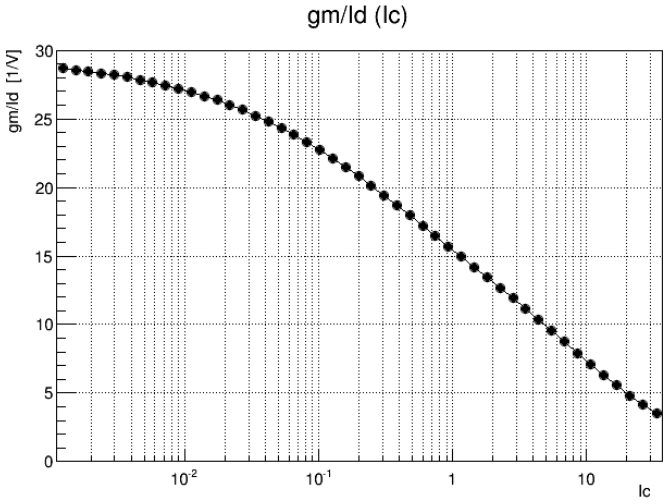

Figure 2.7: g_m/I_D as a function of I_C

2.4 Scaling techniques

Starting from the beginning of the 1970s, the minimum available size of MOS devices have been reduced quickly, thanks to the so-called "scaling techniques". These procedures are not easy, as it has a significant impact on the electrical parameters of the device. Usually the scaling is optimized to have smaller and faster digital gates with reduced cost and power consumption per function and its impact on analog circuits is not always beneficial [29]. A short overview of the main scaling technique is given in this paragraph.
A first method to scale MOS transistor sizes is the so-called "constant electrical field scaling". In this ideal model, voltages and dimensions are scaled by the same factor S, as shown in table 2.1, in order to have the same electrical field in both the original and scaled devices [30]. Constant electric fields ensure the physical integrity of the device and avoids breakdown or other secondary effects. As an example, the basic expression of the current in the saturation region with this scaling becomes:

$$I_{DS,scaled} = \frac{1}{2}\mu_n(SC_{ox})\left(\frac{W/S}{L/S}\right)\left(\frac{V_{GS}}{S} - \frac{V_{TH}}{S}\right)^2 = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2\frac{1}{S} \qquad (2.4.1)$$

It means that the current capability of the devices scales by a factor 1/S [27]. The same applies for the current in linear region. The advantage is given by the reduction of capacitances and power dissipation. As an example, the total channel capacitance becomes:

$$C_{gate} = \frac{W}{S}\frac{L}{S}(SC_{ox}) \qquad (2.4.2)$$

Parameter	Constant Electrical Field Scaling	Fixed-Voltage Scaling	General Scaling
W, L, t_{ox}	1/S	1/S	1/S
V_{DD}, V_{TH}	1/S	1	1/U
C_{ox}	S	S	S
C_{gate}	1/S	1/S	1/S

Table 2.1: Scaling of the main transistor parameters

Concerning the transconductance in saturation region, the following relationship shows that it remains unchanged:

$$g_{m,scaled} = \mu(SC_{ox})\frac{W/S}{L/S}\frac{V_{GS}-V_{TH}}{S} = \mu C_{ox}\frac{W}{L}(V_{GS}-V_{TH}) \tag{2.4.3}$$

Since it can be demonstrated that the width of the depletion region around the drain decreases by a factor S, $\Delta L/L$ remains constant. As a consequence, λ increases by a factor S and the output impedance remains unchanged:

$$r_{0,scaled} = \frac{1}{S\lambda\frac{I_{DS}}{S}} = \frac{1}{\lambda I_{DS}} \tag{2.4.4}$$

As a result, also the intrinsic gain, $g_m r_0$, remains constant.
The reduction of the supply voltage has a big impact in the device behavior. Ideal scaling leads in fact to a reduction of the maximum allowable voltage swings by a factor S, thus reducing the dynamic range. In addition, the power dissipation is lowered:

$$P = \frac{V_{DD}}{S}\frac{I_{DD}}{S} = \frac{V_{DD}I_{DD}}{S^2} \tag{2.4.5}$$

Trying to apply this method between a 350 nm and a 65 nm CMOS technology, the scaling factor would be 5.4. Since the power supply voltage of the 350 nm technology is 3.3 V, in 65 nm it should be scaled to 0.6 V. At the same time, the threshold voltage should decreases from 0.7 V to 0.13 V. Nevertheless, the subthreshold slope is defined by very fundamental physical quantities and does not scale with the shrinking of the size. In principle, when $V_{GS} = 0$ the transistor should be fully off. Unfortunately, in case of a threshold equal to 0.13 V, the current in this configuration is still 2% of the value for $V_{GS} = V_{TH}$. As a consequence, the off-leakage current, defining the static power consumption of digital gates, would be too high. Therefore, the constant electric field scaling is limited by the fact that the subthreshold swing prevents a large reduction of the threshold voltage [29].
The opposite method is the fixed-voltage scaling model. In this case the physical dimensions are scaled down by the factor S, while the power supply and threshold voltages are kept constant. In addition, the electric field inside he device increase by S. However, it also becomes unsustainable since it leads to breakdowns in the gate oxide and in the junction.
As a consequence, a compromise between the two conditions has to be found. The idea is to scale independently the dimensions and the voltages. As a reference, device dimensions are scaled by a factor S, while the voltages are scaled by a factor U smaller than S. In other words, the scaling of power supplies and threshold voltages is slower than the device size one. As an example, the power supply of the 65 nm is 1.2 V, only around a factor 3 lower than the 350 nm case, instead of the original 5.6 factor expected in the constant electric field scaling technique.
A recap of the scaling of the main transistor parameters with the different techniques is given in table 2.1.

2.5 Deep submicron CMOS technologies

Nowadays, the CMOS technologies used in high energy physics applications feature minimum gate lengths which are well below the micron. Therefore they are referred to as "deep submicron technologies". The reduced size of the devices allow to implement complex analog and digital circuits in small silicon areas, making them attractive also for high energy physics applications. Nevertheless, due to the extreme shrinking of their size, the deep submicron devices are characterized by the short-channel effects. In fact, a MOS transistor is defined as a short-channel device if its channel length is on the same order of magnitude as the depletion region thicknesses of the source and drain junctions [35].

2.5.1 Short-channel effects

The short channel effects have to be taken into account for analog design, since they can make the behavior of minimum size transistor very different from the standards. A description of the most important ones is given in the following paragraphs.

Mobility degradation

The vertical electric field rising from the gate voltage contributes to a limitation of the carrier mobility, especially for high V_{GS} values. The electric field, in fact, tends to confine the charge carriers to a narrower region below the silicon-oxide interface, leading to an enhanced charge scattering. As a consequence, a reduction of the surface mobility compared to the bulk one occurs. This effects, called mobility degradation, is modeled by the following equation:

$$\mu_{eff} = \frac{\mu_0}{1 + \eta(V_{GS} - V_{TH})} \tag{2.5.1}$$

in which μ_0 is the bulk mobility and η is an empirical parameter, whose value is around $10^{-7}/t_{ox}\ V^{-1}$. This effect lowers the current capability and the transconductance of the device. In addition, it provokes a deviation of the saturation current from the square law [27]. In fact, inserting the new value of the mobility into equation 2.3.4, it changes as follows:

$$I_{DS} = \frac{1}{2}\frac{\mu C_{ox}}{1 + \eta(V_{GS} - V_{TH})}\frac{W}{L}(V_{GS} - V_{TH})^2 \tag{2.5.2}$$

Assuming $\eta(V_{GS} - V_{TH}) << 1$ it becomes:

$$\begin{aligned} I_{DS} &\simeq \frac{1}{2}\mu_0 C_{ox}\frac{W}{L}[1 - \eta(V_{GS} - V_{TH})](V_{GS} - V_{TH})^2 \\ &\simeq \frac{1}{2}\mu_0 C_{ox}\frac{W}{L}[(V_{GS} - V_{TH})^2 - \eta(V_{GS} - V_{TH})^3] \end{aligned}$$

It is only an approximation, but it reveals the presence of higher harmonics in the drain-source current.

Carrier velocity saturation

The mobility of carriers is also dependent on the lateral electric field in the channel. The carrier velocity is related to the electric field through the following relation:

$$v = \mu\xi \tag{2.5.3}$$

However, when the electrical field along the channel reaches a critical value ξ_c, the velocity of the carriers tends to saturate due to scattering effects, as shown in figure 2.8 [30]. For p-type

silicon, the critical field at which electron saturation occurs is around $1.5 \times 10^6\ V/m$, and the saturation velocity v_{sat} approximately equals $10^5\ m/s$. Therefore, depending on the length, in submicron devices, this condition is easily reached, with a V_{DS} of hundreds of mV.
The consequence of this effect is that the short-channel devices enter the saturation region in advance compared to the long-channel ones, as illustrated in figure 2.2 [30]. Therefore, the constant current is quite lower than the one obtained for large devices. In addition, since an increment in V_{GS} results in a smaller I_{DS} increase, also the transconductance is reduced compared to the square law. In the saturation region, the current under the effect of velocity saturation can be represented as follows [27]:

$$I_{DS} = WC_{ox}v_{sat}\frac{(V_{GS} - V_{TH})^2}{V_{GS} - V_{TH} + 2\frac{v_{sat}L}{\mu_{eff}}} \tag{2.5.4}$$

This relationship leads to some considerations. Firstly, if v_{sat} of L is large, it reduces to the square law current. Therefore, it confirms that this effect is significant only for short devices. Furthermore, if the overdrive voltage is very small the denominator can be approximated with $2v_{sat}L/\mu_{eff}$. Assuming $\mu_{eff} \simeq \mu_0$ the device follows the square law even if L is relatively small. Substituting the value of μ_{eff} this relationship can be further reshaped:

$$\begin{aligned} I_{DS} = WC_{ox}v_{sat}\ &\frac{(V_{GS} - V_{TH})^2}{V_{GS} - V_{TH} + 2\frac{v_{sat}L}{\mu_0}[1 + \eta(V_{GS} - V_{TH})]} = \\ &= WC_{ox}v_{sat}\ \frac{(V_{GS} - V_{TH})^2}{\frac{2v_{sat}L}{\mu_0} + \left(1 + \frac{2v_{sat}L\eta}{\mu_0}\right)(V_{GS} - V_{TH})} = \\ &= \frac{1}{2}\mu_0 C_{ox}\frac{W}{L}\frac{(V_{GS} - V_{TH})^2}{1 + \left(\frac{\mu_0}{2v_{sat}L} + \eta\right)(V_{GS} - V_{TH})} \end{aligned}$$

It is similar to equation 2.5.4, suggesting that the degradation of the mobility with both lateral and vertical fields can be represented adding the terms $\mu_0/2v_{sat}L$ and η. Therefore, the same considerations drawn about mobility degradation can be applied in this case. As an example, it can be shown that the current contains high-order nonlinear terms.

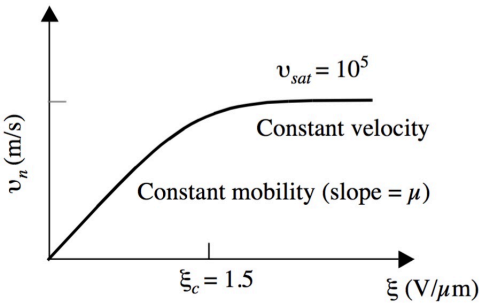

Figure 2.8: Carrier velocity saturation effect [30]

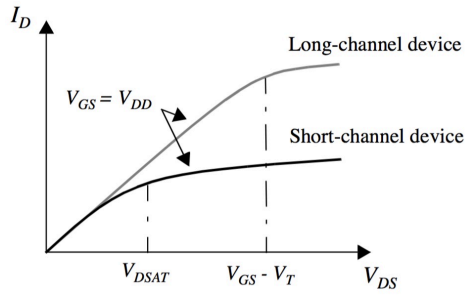

Figure 2.9: $I_D(V_{GS})$ characteristics for long and short-channel devices [30]

Threshold voltage variation

In case of short channel devices, edge effects along their periphery can not be neglected, resulting in significant variations in the overall electric field distribution across the channel. This effect is particularly enhanced if the channel dimensions are not much larger than the combined widths of the depletion regions around source and drain [26]. Edge effects have an important impact on the threshold voltage value. In fact, the depletion regions extend themselves significantly into the channel reducing the immobile charge that must be imaged by the charge on

the gate. Therefore, as figure 2.10 shows, the gate voltage required to form an inversion layer in the charge decreases significantly [27].
A similar effect can be obtained by rising the drain-source voltage, as this increases the width of the drain-junction depletion region [30]. As a result, the threshold decreases with increasing V_{DS}. This effect, called Drain-Induced Barrier Lowering (DIBL) is illustrated in figure 2.11. A significantly high V_{DS} value can even short together the drain and source regions, leading to the so-called punch-through effect. In this case the current increases abruptly leading to possible damage to the transistor. Subsequently, the punch-through has to be avoided and defines and upper limit on the V_{DS} value relevant to the considered device.

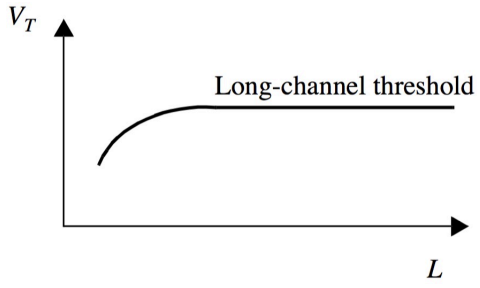

Figure 2.10: Threshold voltage vs channel length [30]

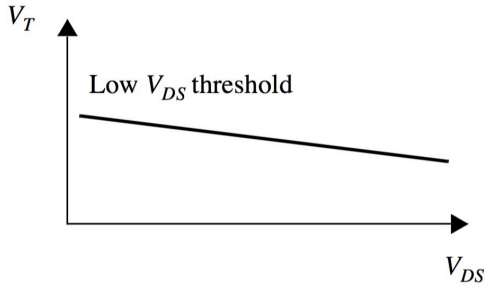

Figure 2.11: Threshold voltage versus V_{DS} [30]

Hot carrier effects

Short-channel transistors featuring a high V_{DS} exhibit high lateral electric fields. Even if the average velocity of carriers saturates at high fields, the instantaneous velocity of carriers tends to increase, in particular in the region near the drain. The latter are named hot carriers. As a consequence, near the drain hot carriers may hit the silicon atoms at high speed, originating impact ionization. Therefore, electron and hole pairs are generated. Electrons are absorbed by the drain and holes by the substrate, giving rise to a finite drain-substrate current. In extreme cases, very high energy carriers can reach the gate terminal, originating a small gate current. The voltage scaling tied to deep submicron technologies reveals itself as an effective solution to keep hot carrier effects under control.

Output impedance variation with V_{DS}

In section 2.3.2 a r_0 constant in the saturation region has been assumed. In reality, r_0 itself has a dependence on V_{DS}. In fact, when V_{DS} increases and the pinch-off point moves towards the source, the rate of the widening of the depletion region around source decreases. As a result, a higher incremental output impedance emerges.
As figure 2.12 shows, in short-channel transistors with a further increase of V_{DS} the DIBL becomes significant. Therefore, the threshold voltage diminishes while the drain current increases and in this region the r_0 remains almost constant. Lastly, at quite high V_{DS} values the impact ionization produces a quite large current, resulting in a lowering of r_0.

Gate leakage current

The reduction of the gate oxide thickness with scaling has increased the tunneling probability of electrons through the oxide. As a consequence, a leakage current is originated. An excessive amount of current through the gate can compromise the performance of the MOS transistor itself. The limit on the SiO_2 thickness is 1.5 nm. The 65 nm process is only marginally beyond this benchmark. As a result, in this technology the gate leakage current has to be taken into

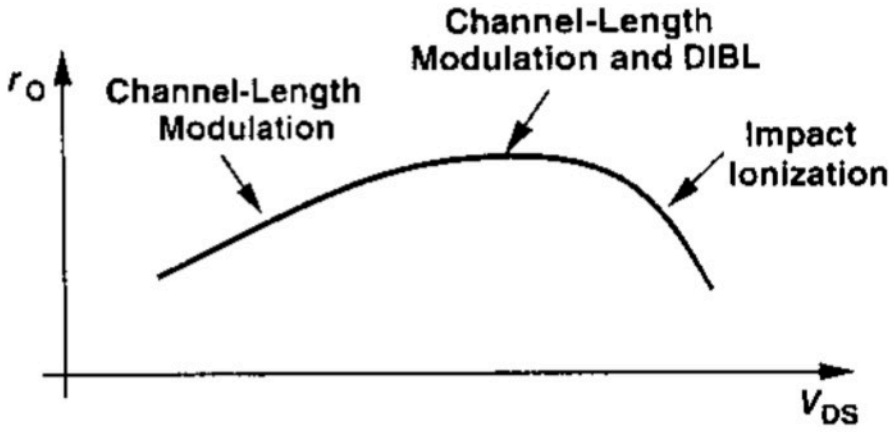

Figure 2.12: Variation of output impedance as a function of V_{DS} [27]

account when many transistors are put in parallel. Subsequent technology nodes, as 45 nm or lower, features new high-k dielectric materials as gate insulators, in order to prevent this problem from becoming dominant. In fact, the high dielectric constant allows to increase the gate capacitance even increasing the oxide thickness, which in turn allows to decrease the gate leakage current contribution. These technologies are not significantly used yet in high energy physics applications due to high cost and necessity of perform a proper characterization under irradiation. However, some preliminary studies on high-k technologies applied to pixel detector readout have been started. An example is provided by the IC-PIX28 chip, presented in [36].

2.5.2 Mismatch effects

In principle two transistors of the same technology and size should have exactly the same performance. In reality, these devices suffer from time-independent random variations in some physical quantities due to uncertainties in the different steps of the manufacturing process. This process is called mismatch. As an example, the gate dimensions of CMOS transistors suffer from random and microscopic variations. Therefore, a mismatch between the lengths and widths of devices identically laid out arises. In addition, these transistors shows a mismatch in the threshold voltage value. The latter, in fact, is a function of the doping levels in the channel and in the gate, which vary randomly between devices [27]. Considering for example the expression of the drain current in saturation 2.3.4, the most of the quantities from which it depends (W,L,C_{ox},μ and V_{TH}) are involved by the mismatch mechanism. As a consequence, these fluctuations have to be minimized, especially in the most delicate part of the designs, in order to avoids significant variations between analog front-end channels.
The exact identification of the mechanisms that cause these effects is significantly difficult, since it can depend on the fabrication technology and the layout. Nevertheless, some basic trends can be described in a quite intuitive way. In fact, mismatch effects are strongly dependent on the gate area. Increasing W and L, random variations have a smaller influence in percentage. A rigorous mathematical observation together with measurements has been performed in [37] [38], stating that:

$$\sigma_{V_{TH}} = \frac{A_{VTH}}{\sqrt{WL}} \tag{2.5.5}$$

$$\sigma\left(\mu C_{ox}\frac{W}{L}\right) = \frac{A_k}{\sqrt{WL}} \tag{2.5.6}$$

A_{VTH} and A_k are the proportionality factors. A_{VTH} is given by the following relationship [39]:

$$A_{VTH} = \frac{qt_{ox}\sqrt{2Nt_{depl}}}{\varepsilon_0\varepsilon_{ox}} \tag{2.5.7}$$

The most important contribution is the number of active doping atoms in the depletion layer ($N = N_A + N_D$). In addition, A_{VTH} scales down with the oxide thickness t_{ox}. On the other hand, transistor scaling forces the doping level under the gate to increase. As a result, the depletion width (t_{depl}) reduces with $N^{-0.5}$. It means that the overall dependence on the doping levels is $N^{0.25}$.
Due to the inverse proportionality with area, mismatch effects are particularly relevant in deep submicron technologies. As an example, even if a 65 nm technology sees an improvement of A_{VTH} compared to a larger one, the minimum size transistor has a significantly smaller area, thus increasing the overall threshold voltage mismatch. Analog front-ends have therefore to be carefully designed in order to minimize these effects in the critical areas of the circuit. Figure 2.13 shows a simulation performed with the 65 nm technology used for the design presented in this work. The threshold voltage dispersion as a function of the device area confirms the expected inverse proportionality.
A detailed description of how the mismatch affects the performance of circuits will be provided in the next chapters.

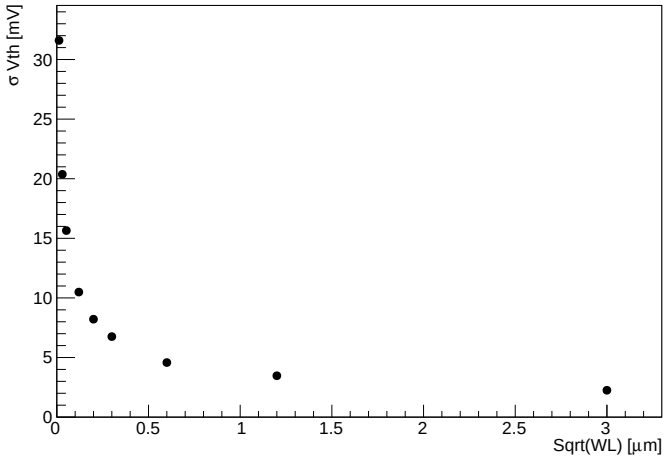

Figure 2.13: $\sigma_{V_{TH}}$ as a function of $\sqrt{WL}$ for a 65nm CMOS technology

2.6 Radiation damage on MOS transistors

As explained in chapter 1, radiation induces a damage in silicon. As a consequence, together with the sensors, also the readout chip suffers a performance degradation due to particle radiation. Regarding CMOS processes, two main groups of effects can show up:

- **Cumulative effects**: they are gradual processes taking place during the whole lifetime of the chip;

- **Single Event Effects (SEE)**: they are due to the energy deposited by a single particle crossing the device. This kind of process can happen at any moment, with a probability defines through a cross section.

2.6.1 Cumulative effects

Concerning the cumulative effects, MOS transistors are sensitive to the TID progressively deposited by particles in the material which constitutes the electronics devices. Common units to measure it are the Gray (Gy) and the rad (1 Gy = 100 rad). A scheme of the cumulative effects which influence the CMOS device behavior is shown in figure 2.14. The main mechanism is the energy deposition in the gate SiO_2. Here the energy is transferred from high energy photons or charged particles through ionization mechanism to generate electron-hole pairs [40]. The energy required to create one pair in the silicon oxide is (17 ± 1) eV. After their creation, electrons and holes start to transport within the oxide. A fraction of them experiences recombination, but this process can happen only during a very short interval of time. In fact in SiO_2 the mobility of electrons at room temperature is around 20 cm^2 V^{-1} s^{-1}, while for holes it is in the interval between 10^{-4} and 10^{-11} cm^2 V^{-1} s^{-1}. As a consequence, electrons quickly move outside the gate oxide, which is quite thin. Holes, instead, are almost fixed in the oxide and are trapped in defect centers in the oxide. This process can also activate defects at the silicon-silicon oxide interface, called interface states.
From the point of view of the gate oxide of MOS transistors, trapped charges screen or enhance, depending on the polarity of the transistor, the gate electric field, leading to a threshold voltage shift. In addition, considering the STI oxides, the trapped holes can attract an image charge in the semiconductor which results in an inversion of the interface. As a consequence, leakage paths between transistor terminals can be opened. The maximum value of leakage current for MOS devices is defined as the drain-source current for $V_{GS} = 0$. The parasitic lateral transistor responsible of this mechanism is illustrated in figure 2.15. The latter issue affects only NMOS devices.
The interface states may instead trap charges from the channel, giving rise not only to a threshold voltage shift but also to a change of the carriers mobility in the channel. In NMOS devices the threshold voltage shifts caused by the trapped holes goes in the opposite direction compared to the one due to interface states. In fact, the former tend to decrease V_{TH} while the latter tend to increase it. For PMOS devices, instead, both effects contribute to an increase of the threshold voltage in absolute value [41].
The dynamic of the two effects is quite different. Holes are in fact trapped quite quickly and can be detrapped by thermal energy through a process called annealing. On the other hand, interface states experience a very slow formation and do not anneal for temperature below 400 °C. In addition, these two effects are very sensitive to the voltages applied to the MOS terminals, since the electron-hole recombination probability is lowered by an increased electric field. Considering the single transistor configuration, the worst case for the NMOS device is $V_{GS} = V_{DD}$ and $V_{DS} = 0$. Regarding the PMOS device it is instead the case in which all the terminals are grounded. Concerning complex circuits featuring many transistors of both flavors, the common choice is to perform TID tests under working conditions. In fact, also in this case a chip irradiation with no bias leads to a significant reduction of the damage and likely an underestimation of the effects that the chip would experience into the experiment during data taking [41].

Another cumulative effect is displacement damage. It consists in the displacement of an atom nucleus in the lattice due to coulombic or nuclear interactions by the interacting particles, creating defects. Nevertheless, in silicon displacement damage results mainly in a reduction of the minority carrier lifetime. Therefore, this phenomenon has little influence on CMOS devices since their current conduction is based on the majority carriers moving in the channel. It is instead relevant in bipolar transistors.

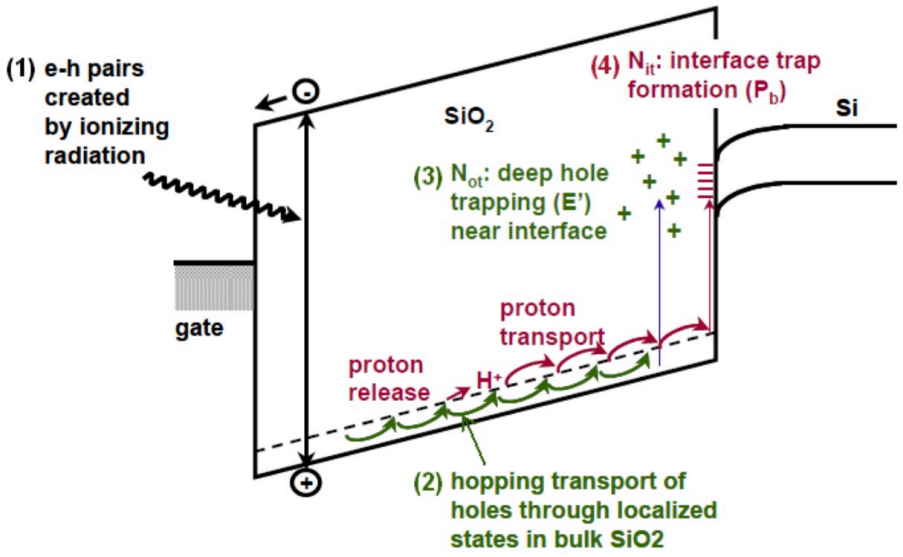

Figure 2.14: Energy band diagram containing the cumulative effects in a MOS device with a positive bias at the gate [40]

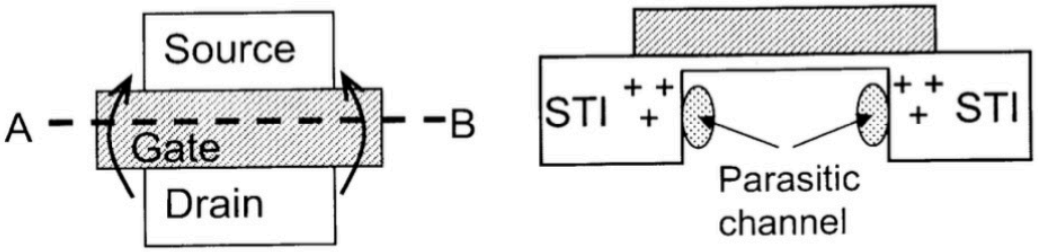

Figure 2.15: Top (left) and lateral (right) view of a MOS transistor representing the radiation-induced leakage paths [41] © [2005] IEEE

2.6.2 Single Event Effects

Single Event Effects are very localized event which can cause a transient, static or permanent damage in the device. Transient errors are the most frequent ones and usually affect digital combinational logic, generating asynchronous signals which propagate through the circuit during one clock cycle. They are therefore called Single Event Upsets (SEU). Given their transient nature, a limited amount of SEU events can be tolerated, based on the system specifications. The most dangerous effects for both analog and digital circuitry are the permanent damages, resulting in destructive events. Concerning CMOS devices, they can be particularly affected by Single Event Latchup (SEL). The basic principles of latchup have been discussed in section 2.1. This effect can also be initiated by the energy deposited by ionizing particles, if it does not recombine. It, in fact, flows to the power supplies to the well and substrate contacts. In case the resistance along this path is high, it results in a large voltage drop, making the local voltage quite different from V_{DD} or ground. The junctions can therefore be forward biased, starting the positive feedback structure which results in the latchup effect. As a consequence, circuits designed for high radiation environment should feature an even higher number of substrate/nwell contacts and guardrings, having to minimize an increased probability of latchup effects compared to normal devices.

2.6.3 Radiation tolerance of deep submicron CMOS technologies

The reduction of transistor sizes has a strong impact on the radiation hardness of CMOS devices. In fact, the channel region is more tolerant to TID thanks to the important reduction of the oxide thickness and the increased doping levels in both channel and bulk areas. In particular the small t_{ox} reduces the number of trapped holes. In addition, a significant reduction of interface states is observed. It appears to be caused by the tunneling of radiation-induced holes outside the oxide before the interface states creation process can start [42].
Nevertheless, the STI structures do not scale down at the same level. As a consequence, macroscopic effects such as source-drain or inter-diffusion leakage currents are still present, limiting the radiation tolerance of CMOS circuits. The leakage current is small compared to the strong inversion saturation current, but can be significant for devices operating in weak inversion. In addition, it can be a limit for NMOS transistors operating as switches, because they will never turn completely off [43]. Figure 2.16 shows the trends of leakage current with radiation for 130 nm devices. It shows that this quantity is not significantly dependent on the width and length of the transistor. In addition, all curves feature a peak in the 1-6 Mrad region. It happens because radiation-induced charges are quickly trapped in the STI oxide at the transistor edge, building up an electric field which opens the inversion channel through the source-drain leakage path, and this phenomenon can only be observed when the current in the parasitic lateral transistor is larger than the current in the main transistor at the same V_{GS}, which happens after around 200 krad. Nevertheless, the negative charge in interface states, which increases slowly, competes with the oxide trapped charge in NMOS transistors: therefore at some Mrad of irradiation level, a rebound effect arises. With further increase of the radiation levels, interface states effects become dominant, limiting the leakage current.

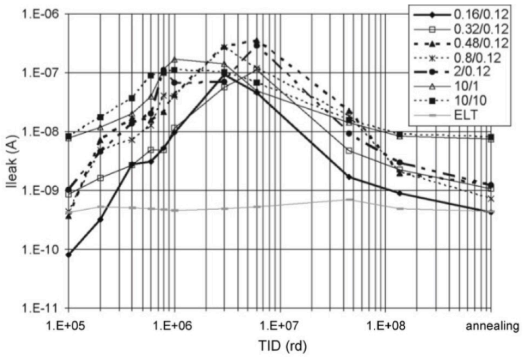

Figure 2.16: Leakage current vs radiation for different sizes of NMOS devices [41] © [2005] IEEE

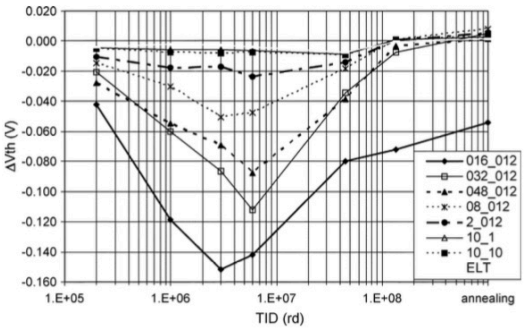

Figure 2.17: Threshold voltage vs radiation for different sizes of NMOS devices [41] © [2005] IEEE

The threshold voltage variation with radiation for different NMOS devices is presented in figure 2.17. It gives evidence of the so-called Radiation-Induced Narrow Channel Effect (RINCE) [43]. In narrow channel devices in fact the charge balance at the transistor edges has also an influence on the electric field of the main transistor, resulting in a decrease of V_{TH} with the gate width. In presence of positive charges in the STI oxide narrow channel effects decrease the threshold of sufficiently narrow NMOS transistors, while they increase it in absolute value for PMOS transistors.
In order to improve the radiation tolerance of an integrated circuit, it is possible to use a specific technique in the transistor layout. In place of the classical design, shown in the left part of figure 2.18, the Enclosed Layout Transistor (ELT) approach showed on the right is adopted [44]. In this kind the gate has an annular shape. This approach forces in fact all the drain-source current to pass underneath the gate oxide, eliminating the leakage paths present in the standard configuration. As shown in figure 2.16, the ELT technique is very effective. In fact,

no significant variations of the leakage current with radiation is observed. As a consequence, at least the most critical transistors, like switches, should be designed as ELT devices. In addition, also the inter-transistor leakage caused by the formation of an inversion layer in the p-type substrate can be limited through a proper layout. In fact, a rise of the doping level of the p-type substrate increases also the threshold for inversion to a very high level, making the positive charge generated in the STI oxide too small to invert the silicon at the oxide interface. This goal is achieved by inserting an uninterrupted p^+ guard ring around the device, separating the n^+ implants from each other [45].

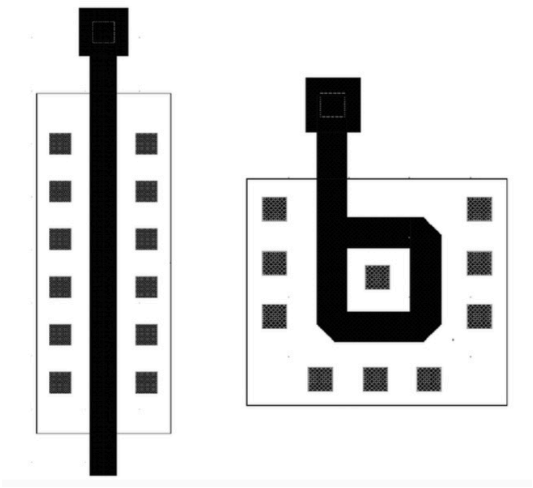

Figure 2.18: Regular gate layout (left) and enclosed layout (right) of a MOS transistor [44]

Regarding ELT transistors, it is important to find a way to relate its dimensions with the standard device ones. In other words, it is important to find an "effective" W/L ratio in order to properly design the ELT device. Based on the scheme illustrated in figure 2.19, some considerations can be drawn [45]. The assumption of square equipotential lines under the gate allows to derive the drain-source current as for a linear transistor. In addition, assuming that the depletion depth is constant underneath the channel and that the gradual channel approximation is valid, it is possible to write:

$$I_{DS} = 8x\mu C_{ox}(V_{GS} - V(x) - V_{TH})\frac{dV(x)}{dx} \tag{2.6.1}$$

In integral form it becomes:

$$I_{DS}\int_{W_1/2}^{W_2/2}\frac{dx}{x} = 8\mu C_{ox}\int_0^{V_{DS}}(V_{GS} - V(x) - V_{TH})dV_x \tag{2.6.2}$$

The integration limits come from the fact that moving from W_1 to W_2 the voltage rises from 0, at the source, to V_{DS} at the drain. Solving therefore the equation it represent the linear region configuration and becomes:

$$I_{DS}\ ln\left(\frac{W_2}{W_1}\right) = 8\mu C_{ox}[(V_{GS} - V_{TH})V_{DS} - \frac{V_{DS}^2}{2}] \tag{2.6.3}$$

Comparing it with the expression of the current in the linear region for the standard device it is possible to write:

$$\left(\frac{W}{L}\right)_{eff} = \frac{8}{ln(W2/W1)} \tag{2.6.4}$$

It represents the effective aspect ratio of an ELT transistor. It has a logarithmic dependence on the device dimensions. As a result, if a very small effective aspect ratio is required, the ELT device has a very large size. If the saturation condition $V_{DS} > (V_{GS} - V_{TH})$ is considered, the same result is obtained.

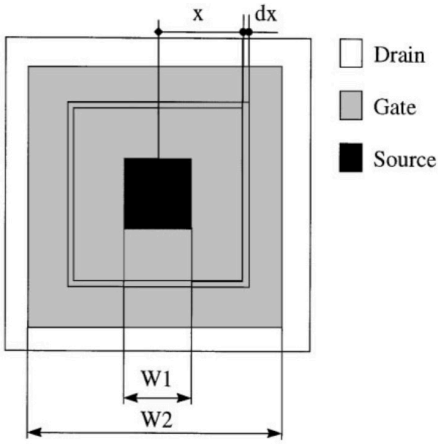

Figure 2.19: Schematic view of a ELT transistor [45]

CMOS 65nm technology

This paragraph is dedicated to the studies on radiation tolerance of the 65nm technology used for the analog design presented in this work. Inside the RD53 collaboration, an important effort in the characterization of the main parameters of single transistors has been performed up to the 1 Grad level. The irradiation campaign has been mainly performed through a X-ray machine at different temperatures [46]. Firstly, room temperature data have been taken. Concerning the minimum size NMOS transistor (i.e. with $W/L = 120n/60n$), figure 2.20 shows that the drain current experiences severe degradation for radiation levels beyond 100 Mrad, decreasing by almost 80% at 1 Grad compared to the pre-irradiation value. Also the transconductance is significantly degraded, with its peak moving to higher V_{GS} values for high-irradiation levels.

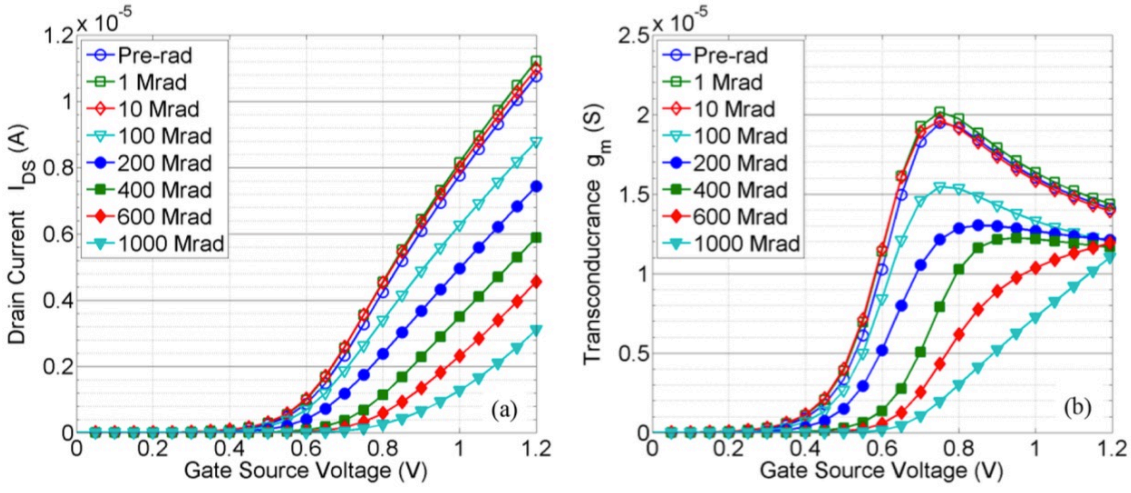

Figure 2.20: NMOS characteristics for the minimum size NMOS transistor biased at $V_{DS} = 50\ mV$ at room temperature [46]

Figure 2.21(a) shows the variation the maximum drain current as a function of radiation levels for differently-sized devices. It shows that this parameter has an important degradation beyond

100 Mrad for all the devices, including the ELT transistors. Looking at the threshold voltage variations presented in figure 2.21(b), it is possible to distinguish two regions. Below 100 Mrad, in fact, no variation is observed, apart from the narrowest device which experiences a small decrease of V_{TH}, resulting in a correspondent increase of the current. Up to this level, therefore, a significant dependence on the transistor width is observed. Beyond 100 Mrad, instead, a sharp increase of the threshold voltage takes place. It tends to turn off the NMOS device, explaining the reduction of the current. This effect has little dependence on the transistor width.

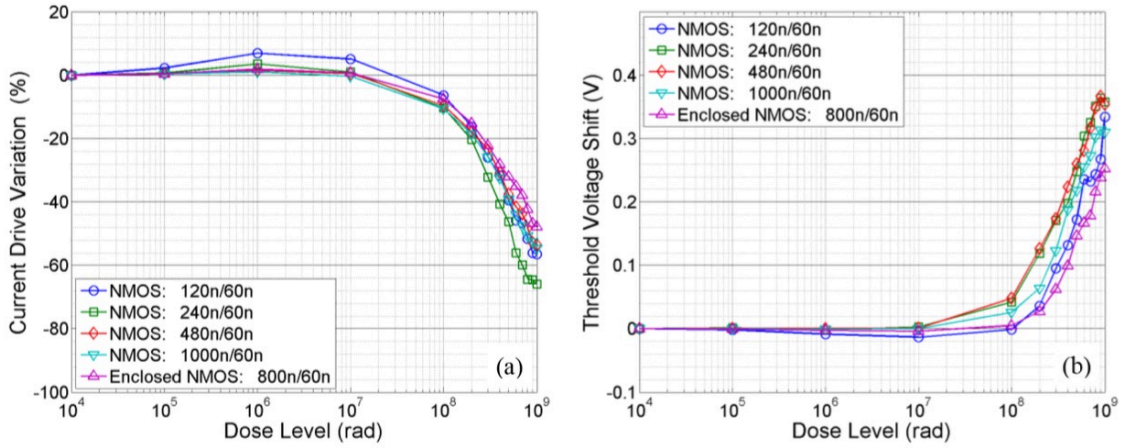

Figure 2.21: NMOS current drive variation for $V_{DS} = 1.2\ V$ (a) and threshold voltage shift for $V_{DS} = 50\ mV$ (b) at room temperature [46]

Regarding PMOS transistors, these effects appear to be enhanced. As figure 2.22(a) shows, for narrow devices the current degradation reaches 100% at 1 Grad, while it is less significant for larger transistors and even more for ELTs. Therefore, for PMOS devices this effect is strongly related with the gate width. Figure 2.22 (b) shows the threshold voltage variations, but not for the smallest devices since they were almost turned off, not allowing the extraction of this parameters.

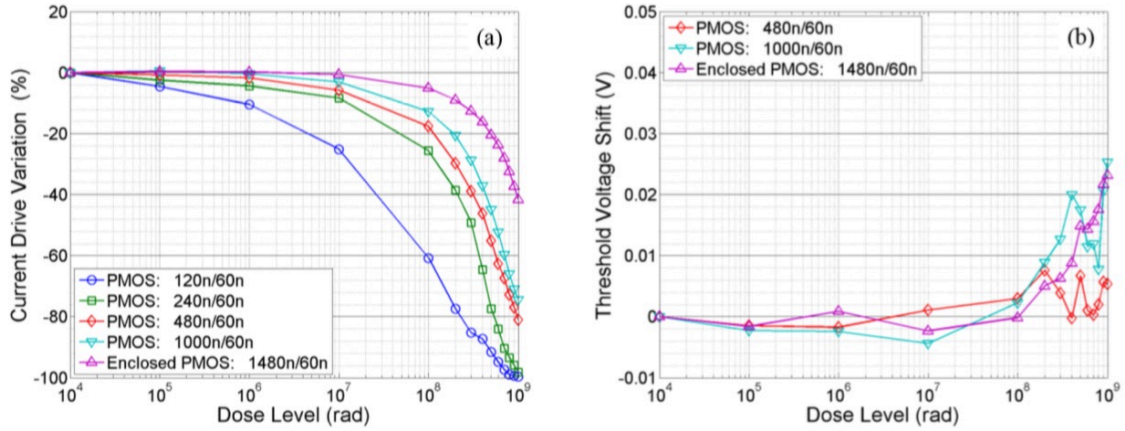

Figure 2.22: PMOS current drive variation for $V_{DS} = 1.2\ V$ (a) and threshold voltage shift for $V_{DS} = 50mV$ (b) at room temperature [46]

Subsequently, since the CMS tracker detector is expected to operate at low temperatures to minimize sensor leakage currents, the devices have been irradiated at around -15 °C. A comparison of the current degradation with the room temperature configuration is illustrated in figure 2.23 for both minimum size NMOS and PMOS devices. At 1 Grad, the former exhibit

a 20% improvement, while the latter, which were completely off, experience a 60% degradation. Low temperature operation is therefore quite helpful also for the readout chip.

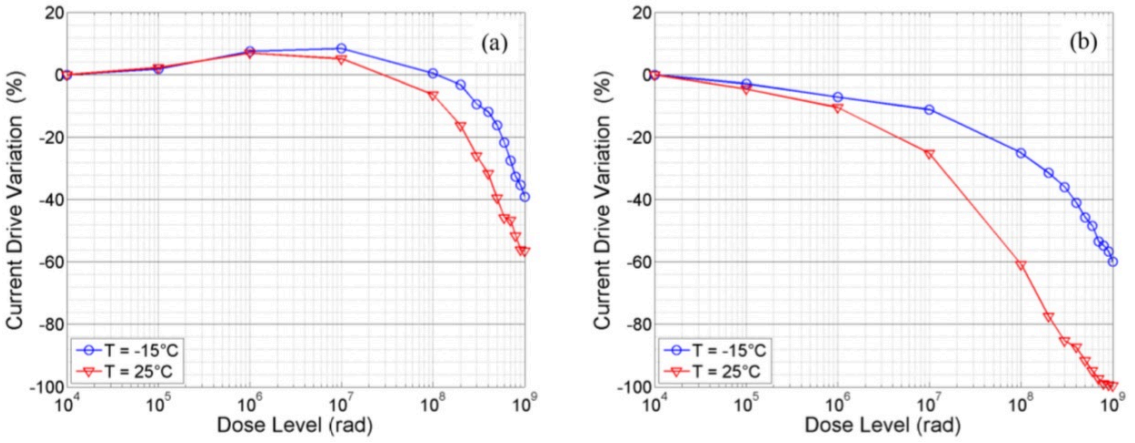

Figure 2.23: Current drive variation for $V_{DS} = 1.2V$ versus TID for a $120n/60n$ NMOS (a) and a $120n/60n$ PMOS (b) [46]

After irradiation, the devices have been annealed at different temperatures (-15, 25 and 100 °C). NMOS transistors have not shown any sign of recovery. PMOS devices, instead, are characterized by a small recovery at -15 °C, a steady behavior at room temperature and a further degradation at 100 °C, due to a sharp increase of the threshold voltage. As a result, the annealing process looks to be little effective on the single device and should not be performed at high temperatures.
Keeping in mind these considerations, some precautions have to be taken in analog design. In particular, it is advisable to avoid the usage of minimum size PMOS transistors, given that a reasonable increase in the gate width reduces significantly the radiation damage.

Chapter 3

Front-end amplifiers

In this chapter an overview of the main features of analog front-ends for the readout of silicon sensors is given. The functionalities of the building blocks like preamplifier, shaper and discriminator are discussed, together with key parameters like peaking time and noise. The second part of the chapter is dedicated to the main types of analog front-ends: sample-and-hold, binary, counting and ToT architectures. For each of them a practical implementation for high energy physics experiment is described.

3.1 Main aspects in front-end electronics design

The designs of Application Specific Integrated Circuits (ASIC) for silicon sensor applications feature some common aspects. These chips are usually composed of an array of identical channels. Figure 3.1 shows a typical detector channel. The incident radiation is absorbed in the sensor and converted in an electrical signal. Then it is fed into the analog chain, featuring an input amplifier, a shaping filter, an additional signal processing and some memories to store the data [47]. In certain applications, only some of these blocks are used.

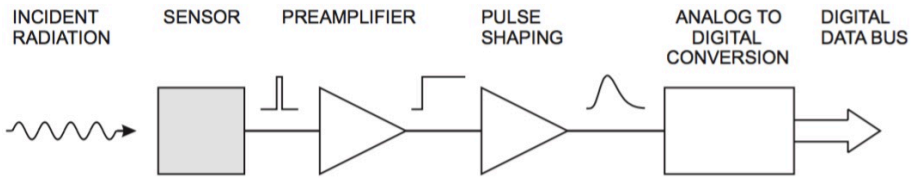

Figure 3.1: Typical sensor-analog front-end system [47]

The analog front-end design process is performed taking into account the tasks that it has to carry out: it should be able to acquire the electrical signal from the sensor and then adapt it to a specific measurement depending on the application (for example position, time or energy measurement).
Since it has an impact on the analog front-end implementation, it is crucial to model the electrical behavior of the sensor. A typical model is shown in figure 3.2. It contains three elements:

- The sensor **leakage current**, I_{leak}, represented by a DC generator. Obviously, it can be considered constant only for a fixed level of irradiation, otherwise it changes with time. It can have an influence on the preamplifier behavior. As a consequence, leakage current compensation circuits are often implemented in front-end designs;
- The **total sensor capacitance**, C_{det}, which is, as illustrated in section 1.6.1, the sum of the backside and interpixel contributions. It has a significant influence on noise;

- The **sensor output signal**, modeled as a time-dependent current source I_{pulse}. Its exact shape depends on many factors like the position of charge deposition, the sensor material properties, the bias voltage and the pixel geometry. The polarity of the signal is determined by the type of charge carriers collected on the pixel [19].

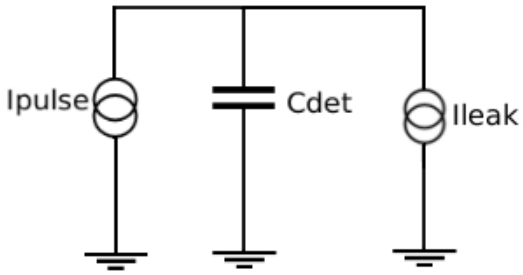

Figure 3.2: Equivalent circuit of a silicon pixel sensor

The next paragraphs contain instead a description of the main building blocks and design parameters which are in common between almost all the analog front-end designs.

3.1.1 Preamplifier gain

In analog front-ends for particle sensors, a key role is played by the first amplification stage, usually referred to as "preamplifier". In fact, it has to convert the signal coming from the sensor, which is usually quite small, into a signal able to drive the following part of the readout chain. It is commonly designed as a transimpedance amplifier connecting an appropriate network in the feedback path of a high gain voltage amplifier. This choice derives from the fact that the sensor signal is in form of a current and should be converted into a voltage.
Some first considerations can be drawn from an ideal preamplifier model, shown in figure 3.3. The transfer function can be studied making some assumptions:

- The input signal can be modeled with a Dirac delta.
- Ideal core amplifiers, therefore characterized by infinite gain and bandwidth.
- The preamplifier is an ideal integrator. As a consequence, the feedback resistor plays only the role of establishing the correct DC bias point for the input transistor, and can be chosen large enough not to be involved in the signal processing.

As a consequence it is possible to write:

$$V_{out} = \frac{1}{C_f} \int I_{in}(t)dt \tag{3.1.1}$$

Inserting the delta-like input:

$$V_{out} = \frac{1}{C_f} \int Q_{in}\delta(t)dt \tag{3.1.2}$$

Integrating the delta it becomes:

$$V_{out} = \frac{Q_{in}}{C_f} u(t) \tag{3.1.3}$$

Therefore, the voltage output signal for an ideal amplifier is described by a step function. Nevertheless, in real amplifiers the rise time is not negligible. In addition, this relationship

shows that the output amplitude is inversely proportional to the feedback capacitance. In the frequency domain, the relationship becomes:

$$V_{out} = \frac{Q_{in}}{C_f}\frac{1}{s} \tag{3.1.4}$$

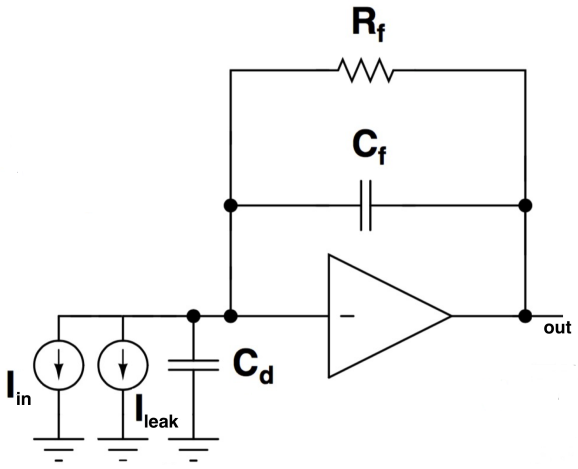

Figure 3.3: Scheme of an ideal preamplifier with a sensor at the input [29]

Real amplifiers are characterized by a finite gain A_0. As a consequence, the input node is not anymore a virtual ground. In the frequency domain the node equation becomes [29]:

$$-I_{in}(s) + V_{in}sC_{det} + (V_{in} - V_{out})sC_f = 0 \tag{3.1.5}$$

Knowing that

$$V_{in} = -\frac{V_{out}}{A_0} \rightarrow V_{out} = -V_{in}A_0 \tag{3.1.6}$$

the previous relationship becomes:

$$-I_{in}(s) + V_{in}sC_{det} + V_{in}sC_f + V_{in}sA_0C_f = 0 \tag{3.1.7}$$

Assuming still a delta-like input, in the frequency domain the Laplace transform of the delta is 1. Therefore, I_{in} it is simply equal to Q_{in}:

$$-Q_{in} + V_{in}sC_{det} + V_{in}sC_f + V_{in}sA_0C_f = 0 \tag{3.1.8}$$

Extracting V_{in} it becomes:

$$V_{in} = \frac{Q_{in}}{s[C_{det} + C_f(1 + A_0)]} \tag{3.1.9}$$

This value represents the small residual voltage which remains at the input due to the finite gain of the preamplifier. It shows that the higher the gain the smaller this contribution is. It is then possible to define the effective input capacitance as follows:

$$C_{eff} = (1 + A_0)C_f \tag{3.1.10}$$

It acts in parallel to the detector capacitance C_{det}. Using again equation 3.1.6 it is possible to find the output voltage step for a finite gain amplifier [19]:

$$V_{out} = -\frac{A_0Q_{in}}{C_f}\frac{1}{s[\frac{C_{det}}{C_f} + 1 + A_0]} \rightarrow -\frac{Q_{in}}{C_f}\frac{1}{s[1 + \frac{1}{A_0} + \frac{C_{det}}{C_fA_0}]} \tag{3.1.11}$$

As a result, a significant portion of the amplitude is lost in case C_{in} is close to $A_0 C_f$. In order to minimize losses, it is therefore important to have a preamplifier with high gain and in which C_{eff} is significantly larger than the detector capacitance C_{det}. In this way also the input residual voltage is kept low, reducing the cross coupling between pixels caused by the interpixel component of the input capacitance.
In addition to these considerations, it has to be underlined that the feedback network has an influence also in signal processing. This aspect can be understood by analyzing the scheme presented in Figure 3.4 which shows a typical transimpedance amplifier with an ideal feedback. The main characteristics of this system are: a main amplifier providing the highest possible gain, a feedback network which senses the output and takes a portion of it back to the input and a subtracting node. The latter, calculating the difference between the input and feedback signals, generates the so-called "error-signal", which is used to drive the amplifier [29].

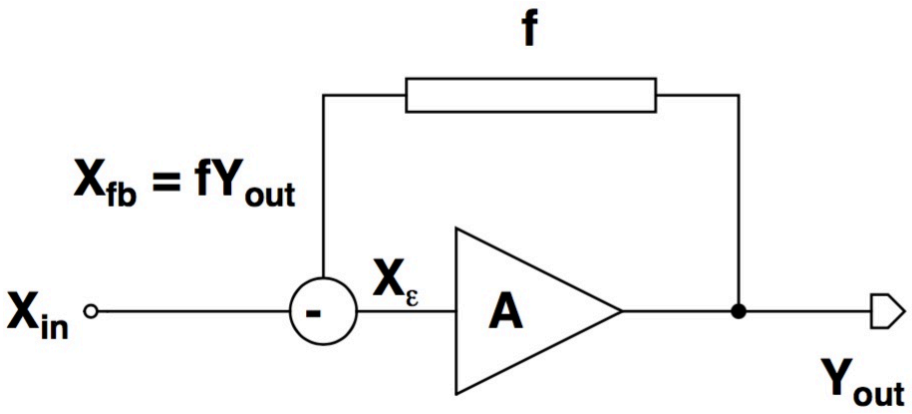

Figure 3.4: Amplifier with feedback [29]

A quantity which is often reported like a figure of merit of a front-end design is the open-loop gain, A. In other words, it is the preamplifier gain in case no feedback is applied. In a feedback system, the amplifier output is sensed by the feedback network, which sends back to the input a portion of it, fY_{out}. Therefore, the error signal is given by:

$$X_\varepsilon = X_{in} - fY_{out} \tag{3.1.12}$$

Since this is the signal which drives the amplifier, the output signal Y_{out} can be expressed as follows:

$$Y_{out} = AX_\varepsilon = A(X_{in} - X_{fb}) = A(X_{in} - fY_{out}) \tag{3.1.13}$$

Starting from this expression it is possible to define the amplifier closed-loop gain, given by:

$$A_{CL} = \frac{Y_{out}}{X_{in}} = \frac{A}{1 + Af} \tag{3.1.14}$$

Some additional considerations can be drawn. If $Af >> 1$ the closed loop gain can be approximated with $1/f$. Therefore, it is defined only by the parameters of the feedback network. Since $f \leq 1$ the feedback network behaves like an attenuator and can be built with quite simple circuits [29]. In addition, considering a variation of the open loop gain ΔA it is possible to write:

$$A_{CL} = \frac{A + \Delta A}{1 + f(A + \Delta A)} \simeq \frac{A}{1 + Af} + \frac{\Delta A}{1 + Af} \tag{3.1.15}$$

As a consequence is it not crucial to control precisely the value of the preamplifier gain, it is enough to make is large.
The gain is commonly expressed in mV/fC, since the ratio between the output amplitude and

the input charge is considered. An alternative way, very often used in HEP applications, is to use mV/ke$^-$. In fact, Q_{in} can be expressed in units of electron charge keeping in mind the conversion 1 fC = 6250 electrons. In addition, it is very important to known the range of input signals emerging from the detector. In fact, if a linear output is required the preamplifier should be designed in order to avoid saturation for the largest values. This goal is reached by tuning the gain but also choosing an adequate DC baseline voltage at the output, leaving enough headroom in the whole range. In this context, it is also crucial to know the input signal polarity. Given that the preamplifier shows an inverting behavior, in fact, the DC baseline should be set in order to leave the larger margin on the right side, negative or positive depending on the application.

3.1.2 Shaping stage

As shown in figure 3.1, the preamplifier is followed by a pass-band filter, usually called shaper. It is often used in analog front-ends because it helps in the reduction of high and low-frequency noise components. The easiest implementation, illustrated in figure 3.5, is composed of a high-pass CR filter followed by a low-pass RC one.

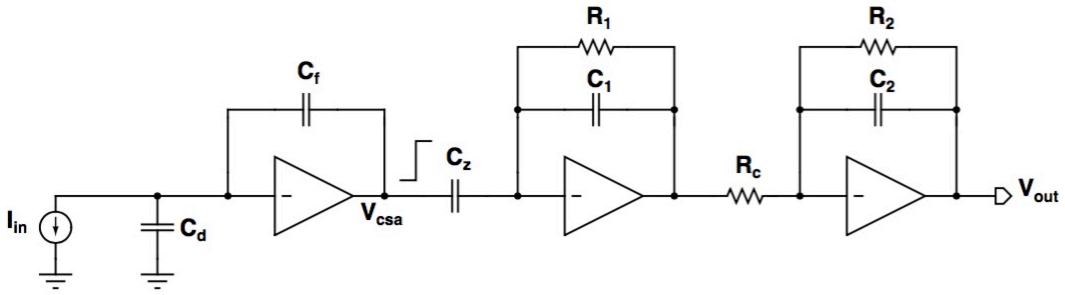

Figure 3.5: Implementation of a CR-RC shaper [29]

The transfer function of this circuit should be analyzed in detail [29]. The Charge Sensitive Amplifier (CSA) output voltage is again converted into a current flowing in C_z:

$$I_{C_z}(s) = \frac{Q_{in}}{sC_f} sC_z = I_{in}(s)\frac{C_z}{C_f} \tag{3.1.16}$$

Then the second stage is again a transimpedance amplifier characterized by a complex impedance in its feedback path. As a result, the current is converted into a voltage at the output of the stage, which has the following expression:

$$V_{out2} = Q_{in}\frac{C_z}{C_f}\frac{R_1}{(1+sC_1R_1)} \tag{3.1.17}$$

The resistor R_c converts V_{out2} back to a current while the last stage provides a voltage as an output. Therefore, the final transfer function is given by:

$$V_{out} = Q_{in}\frac{C_z}{C_f}\frac{R_1}{1+s\tau_1}\frac{1}{R_c}\frac{R_2}{1+s\tau_2} \tag{3.1.18}$$

The implementation of a CR-RC requires the matching of the two time constants. Therefore, the transfer function changes as follows:

$$V_{out} = Q_{in}\frac{C_z}{C_f}\frac{R_1R_2}{R_c}\frac{1}{(1+s\tau)^2} \tag{3.1.19}$$

Using the inverse Laplace transform, it can be written in the time domain:

$$V_{out}(t) = Q_{in}\frac{C_z}{C_f}\frac{R_1R_2}{R_c}\frac{1}{\tau}\left(\frac{t}{\tau}\right)e^{-\frac{t}{\tau}} \tag{3.1.20}$$

Peaking time

The value of the CR-RC shaper time constant defines also the peaking time T_P, which is a key figure of merit for an analog front-end. In fact, as it will be shown in detail in the next chapter, it has a strong influence on the noise figure, but also on the speed. It corresponds to the time at which the transfer function reaches its maximum value:

$$\frac{dV_{out}}{dt} = Q_{in}\frac{C_z}{C_f}\frac{R_1R_2}{R_c}\frac{1}{\tau}\left(\frac{1}{\tau}e^{-\frac{t}{\tau}} - \frac{t}{\tau^2}e^{-\frac{t}{\tau}}\right) = 0 \tag{3.1.21}$$

$$1 - \frac{t}{\tau} = 0 \rightarrow t = \tau = T_P \tag{3.1.22}$$

Therefore, the peaking time of a CR-RC shaper corresponds to the time constant of the filters. Using this relationship it is also possible to evaluate the amplitude of the output signal at the peak:

$$V_{out,peak} = V_{out}\Big|_{t=\tau} = Q_{in}\frac{C_z}{C_f}\frac{R_1R_2}{R_c}\frac{1}{\tau}\frac{1}{e} = \frac{Q_{in}}{C_f}\frac{C_z}{C_1}\frac{R_2}{R_c}\frac{1}{e} \tag{3.1.23}$$

in which the conversion $\tau = R_1C_1$ has been used. As a consequence, the gain of the final stage should have a value equal to e to maintain the preamplifier one unchanged. Figure 3.6 represents this configuration. The preamplifier provides a step function with amplitude Q_{in}/C_f as an output. It is then reshaped by the derivator and the integrator stages, designed to maintain the gain.

The peaking time value can span a very large range of values depending on the application. As an example, timing measurements need a very fast architecture, in the order of few nanoseconds. In addition, it is important to keep in mind that only if the preamplifier peaking time is significantly larger than the charge collection time of the sensor the output of the preamplifier is well described by the expression of the transfer function for a Dirac-delta input pulse. Otherwise, the output will be a convolution between the detector signal and the front-end response to the Dirac-delta. As a result, the amplitude of the output signal will be reduced. This effect is called "ballistic deficit".

Often the shaping stage features more than one integrator. In such a case, it is referred to as "CR-RC^n shaper". The number of integrator defines the order of the shaper. The transfer function contains therefore n+1 poles and in the Laplace domain it is equal to:

$$V_{out}(s) = \frac{Q_{in}}{C_f}\frac{C_zR_1}{(1+s\tau)^{n+1}} \tag{3.1.24}$$

while in the time domain it becomes:

$$V_{out}(t) = \frac{Q_{in}}{C_f}\frac{C_zR_1}{n!\,\tau}\left(\frac{t}{\tau}\right)^n e^{-\frac{t}{\tau}} \tag{3.1.25}$$

In this case the peaking time is given by the time at which:

$$\frac{dV_{out}}{dt} = \frac{Q_{in}}{C_f}\frac{C_zR_1}{n!\,\tau}\left[n\left(\frac{t}{\tau}\right)^{n-1}\frac{1}{\tau}e^{-\frac{t}{\tau}} - \left(\frac{t}{\tau}\right)^n\frac{1}{\tau}e^{-\frac{t}{\tau}}\right] = 0 \tag{3.1.26}$$

$$n\left(\frac{t}{\tau}\right) - 1 = 0 \rightarrow t = n\tau = T_P \tag{3.1.27}$$

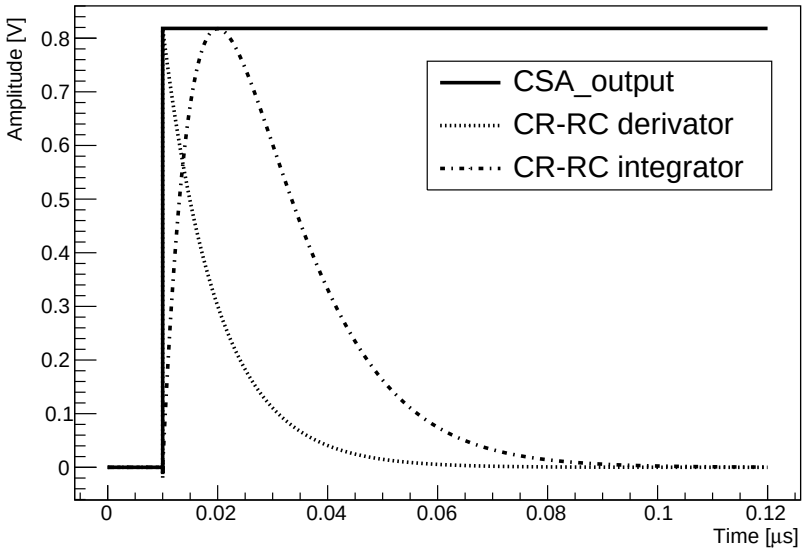

Figure 3.6: Comparison between the CSA, derivator and integrator outputs

As a result, in a CR-RC^n shaper the peaking time is proportional to the number of shapers inserted. Also in this case the value of the amplitude of the output signal corresponding to the peaking time can be found:

$$V_{out,peak} = V_{out}\Big|_{t=n\tau} = \frac{Q_{in}}{C_f}\frac{C_z}{C_1}\frac{1}{n!}n^n e^{-n} \tag{3.1.28}$$

The ratio between the peak amplitudes for two shapers of order n and n-1 leads to some considerations. It is given by:

$$\frac{V_{out,peakn}}{V_{out,peakn-1}} = \frac{n^n}{n(n-1)^{n-1}}\frac{1}{e} \tag{3.1.29}$$

Taking the limit for $n \to \infty$:

$$\lim_{n\to\infty} \frac{n^n}{n(n-1)^{n-1}} = 1 \tag{3.1.30}$$

Therefore, for high order shapers the previous ratio becomes equal to 1. Nevertheless, usually the number of integrators is kept below 10 in order to find the best compromise between shaping and number of active stages, which can be quite area demanding. Figure 3.7 shows how the signal changes with the increase of the order of the CR-RC stage.

3.1.3 Discriminator

The discriminator is the last main block of the front-end chain. The purpose of this stage is to convert the analog information in a digital quantity. It is usually composed of a differential stage which performs a comparison between the analog signal coming from the shaper and a threshold value globally distributed to all channels. The latter is set as low as possible in order to detect very small signals, but keeping an adequate margin with respect to noise, in order to minimize fake hits. The transistor mismatch effects result in a fluctuation of the effective threshold value between pixels. In order to minimize this contribution, compensation technique are included in analog front-end designs. These methods are described in the next chapter.

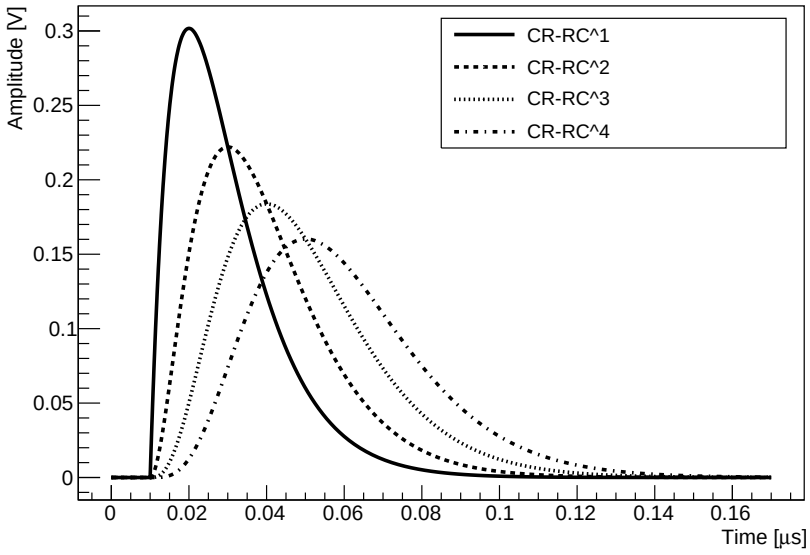

Figure 3.7: Output signal of shapers with different numbers of integrators

3.1.4 Noise

The electronic noise is one of the most important parameter to take into account. In fact, it limits the minimum signal level that can be processed by the analog circuitry at an acceptable level. Since it is a random phenomenon, instantaneous values of noise can not be predicted. As a consequence, the way of including it into circuit analysis is to perform long-time observations used to build a statistical model for the noise. The main predictable aspect of noise is its average power. In fact, the most noise sources exhibit a constant average power. In order to define this quantity in a proper way, it is useful to recall that the average power delivered by a voltage $v(t)$ to a load resistance R_L is [27]:

$$P_{av} = \frac{1}{T}\int_{-T/2}^{T/2} dt\ \frac{v^2(t)}{R_L} \tag{3.1.31}$$

in which T is the period. In order to define P_{av} for a random $v(t)$ voltage, it is necessary to perform the measurement over a long time:

$$P_{av} = \lim_{T\to\infty}\int_{-T/2}^{T/2} dt\ \frac{v^2(t)}{R_L} \tag{3.1.32}$$

In addition, it is usual to simplify this relationship as follows:

$$P_{av} = \lim_{T\to\infty}\int_{-T/2}^{T/2} dt\ v^2(t) \tag{3.1.33}$$

In this model, P_{av} is therefore calculated in V^2 instead of W.
In addition to this first expression, it should be taken into account the frequency spectrum of noise. It is also frequently called power spectral density (PSD). The PSD $S_v(f)$ of a noise waveform $v(t)$ is defined as the average power carried by $v(t)$ in a one-hertz bandwidth around f. It is expressed in V^2/Hz but it is often represented with its square root, therefore expressed in $V/\sqrt{Hz}$. The simplest type of noise PSD is the so-called "white noise". In this case the PSD has the same value of all the frequency values. As a result, the white noise spectrum is

independent from the frequency.
Electronic devices are subject to different types of noise. A short overview of the most important ones is provided in the next paragraphs.

Thermal noise

The random motion of electrons in conductors leads to fluctuations in the voltage measured across it even with an average current equal to zero. This effect, called "thermal noise", is therefore proportional to the absolute temperature.

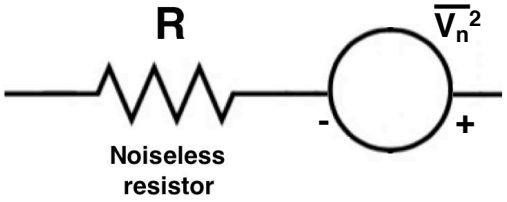

Figure 3.8: Thermal noise of a resistor

Figure 3.8 shows that for a resistor this contribution can be modeled like a voltage source in series with the resistor. The spectral density is given by:

$$S_v(f) = 4kTR \tag{3.1.34}$$

As a consequence the resistor noise spectrum is white. A dependence on frequency happens only for values beyond 100 THz, which are extremely high for the HEP applications. Therefore, the white noise approximation is valid.
Concerning the MOS transistors, the main thermal noise source is generated in the channel. As shown in figure, considering long channel devices in saturation the noise contribution can be modeled like a current generator connected between the source and drain terminals and characterized by the following thermal density:

$$\overline{I_n^2} = 4kT\gamma g_m \tag{3.1.35}$$

in which γ is the inversion factor defined in chapter 2. Nevertheless, in short channel devices an increase of noise compared to the pure thermal noise predictions is observed. This effect is called excess noise and leads to a modification of the previous expression for short channel transistors:

$$\overline{I_n^2} = 4kT\gamma g_m \alpha_w \tag{3.1.36}$$

α_w is therefore called excess noise factor. The mechanism behind this noise increase is not yet fully understood. Nevertheless, some theories have been developed. One of them suggests that it is due to the fact that, during their transit in a very short channel, the carriers may not experience enough collisions to reach the thermal equilibrium, which is the basic assumption of the thermal noise models.
Also the ohmic section of the transistor slightly contributes to the thermal noise. In fact, gate, source and drain materials are characterized by a finite resistivity. Nevertheless, source and drain components are usually negligible and only the gate one can be relevant [27].

Flicker noise

The flicker noise is characterized by a $1/f$ spectrum, therefore dominates at low frequencies. A number of models have been developed in order to describe this phenomenon. Originally, it has been thought as a surface effect, concerning the interface between gate oxide and channel. The

idea is that since the silicon crystal reaches an end at this interface, many bonds appear giving rise to extra energy states. When the charges move towards the interface, some are trapped and then released, introducing flicker noise in the drain current. Measurements have shown in fact a good correlation with the density interface or near interface oxide traps. Subsequently, another model considering instead mobility fluctuations due to defect scattering as the origin of $1/f$ noise has been developed. Therefore it also assumes a volume origin in place of a surface one. Anyway, what is clear is that flicker noise is a function of the technology and is influenced by the material quality and by the number of defects at both sides of the Si-SiO_2 interface [48]. At circuit level, it is then modeled with a voltage source in series with the gate given by:

$$\overline{V_n^2} = \frac{K_f}{C_{ox}WL}\frac{1}{f} \tag{3.1.37}$$

K_f is a process-dependent constant with a value around $10^{-25}V^2F$. This expression is only a first order approximation, since the exact description of the phenomenon is quite complex. This relationship also suggests that larger devices are less affected by flicker noise. As an example, a comparison of the total spectral density for two NMOS transistors in 65nm technology has been carried out. As figure 3.9 shows, a minimum size transistor with $W/L = 200nm/60nm$ is characterized by a significantly large flicker noise contribution compared to a device with $W/L = 500nm/150nm$. The plot also shows that at frequencies beyond 10 MHz the flicker noise contribution becomes almost negligible and the noise spectrum is dominated by white noise components.

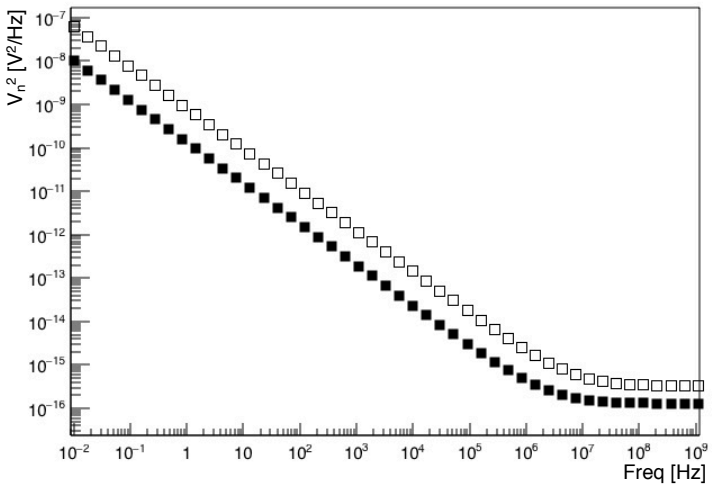

Figure 3.9: Noise spectral density for 65nm technology transistors; the white squares represent a NMOS with $W/L = 200nm/60nm$, the black squares a NMOS with $W/L = 500nm/150nm$

3.1.5 Shot noise

In addition to MOS devices, also the sensor leakage current contributes to the total noise figure. In fact, as illustrated in the model of figure 3.2, it can be represented by a current source in parallel to the front-end input. It is a shot noise contribution. This kind of noise is in fact originated by the fact that in reality a DC current is not a continuous flow, but can be represented as a sum of discrete pulses in time, corresponding to the transfer on an electron through the conductor. Therefore, also shot noise is a random process. Assuming I_{DC} as the mean value of the current and $I(t)$ as the instantaneous value of the current, the shot noise

can be defined as follows [49]:

$$i_{shot} = \sqrt{\overline{(I(t) - I_{DC})^2}} \tag{3.1.38}$$

The bar implies a time average. During a time interval Δt the DC current is given by:

$$I_{DC} = \frac{e\overline{n}}{\Delta t} \tag{3.1.39}$$

in which $\overline{n}$ is the average number of carriers produced in the time Δt. The same expression applies for the instantaneous current:

$$I(t) = \frac{en(t)}{\Delta t} \tag{3.1.40}$$

in which $n(t)$ is the instantaneous number of charged particles emitted in the instant Δt. Since it is a configuration with independent emission events with a constant mean $\overline{n}$ and a variation σ^2, $n(t)$ can be described by a Gaussian distribution:

$$P(n(t)) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(n(t)-\overline{n})^2}{2\sigma^2}} \tag{3.1.41}$$

Assuming $\overline{n} >> 1$ but also that for most of the time there is no charge carrier reaching the detector, it is possible to use a special case of the Gaussian distribution, which is the Poisson distribution, characterized by $\sigma^2 = \overline{n}$. Using this distribution

$$P(n(t)) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(n(t)-\overline{n})^2}{2\overline{n}}} \tag{3.1.42}$$

it is possible to calculate:

$$\overline{(n(t) - \overline{n})^2} \simeq 2\overline{n} \tag{3.1.43}$$

As a result:

$$i_{shot} = \sqrt{\overline{(I(t) - I_{DC})^2}} = \frac{e}{\Delta t}\overline{(n(t) - \overline{n})^2} = \frac{e}{\Delta t}\sqrt{2\overline{n}} \tag{3.1.44}$$

Using equation 3.1.39 it becomes:

$$i_{shot} = \sqrt{\frac{2I_{DC}\ e}{\Delta t}} \rightarrow i^2_{shot} = 2I_{DC}\ e\Delta f \tag{3.1.45}$$

Therefore, recalling the definition of power spectral density which is given for an interval of frequency of 1 Hz, the sensor leakage current is characterized by:

$$i^2_n = 2q_e I_{leak} \tag{3.1.46}$$

As a consequence, this quantity increases during the detector operation, since irradiation results in the increase of the leakage current.

Noise representation in analog front-ends circuits

Based on the previous considerations, an analog front-end is affected by a number of noise sources: the devices it is composed of and the sensor leakage current. Circuit analysis is therefore performed using equivalent noise sources with the purpose of referring to the input the noise of a given device. This choice is made to get rid of the gain dependence on noise. In fact, the input-referred noise and the input signal are both multiplied by the gain during the processing through the circuit. In this way the input-referred noise gives an indication of how much the input signal is corrupted by the circuit noise. In other words, it allows to understand which is the smallest signal that can be detected with an acceptable signal-to-noise

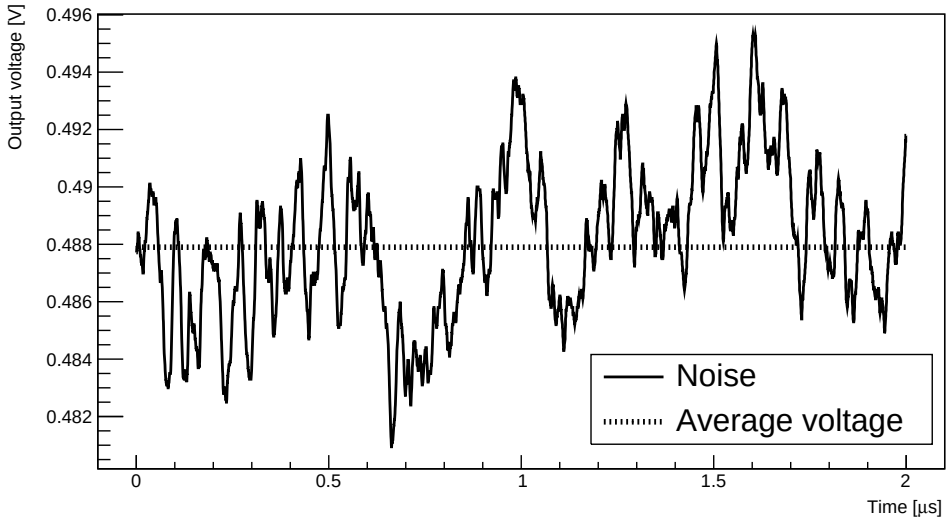

Figure 3.10: Evolution of output noise with time for a typical amplifier superimposed to the average output voltage

ratio (SNR). Input-referred noise is therefore an adequate quantity to perform a fair comparison between different circuits [27]. Figure 3.10 shows the effect of noise on the output baseline of an amplifier, giving evidence of the random nature of this phenomenon.
The procedure consists therefore in measuring the standard deviation of the noise at the output of the system given by

$$V_{Noise} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{n}(V_i - V_{mean})^2} \tag{3.1.47}$$

and then divide it by the gain. The quantity usually adopted to quote the input-referred noise thus obtained is the Equivalent Noise Charge (ENC). If the gain is expressed in mV/ke$^-$ the ENC is then given in number of electrons. The total noise can also be expressed as the sum in quadrature of the different contributions:

$$ENC = \sqrt{ENC_w^2 + ENC_i^2 + ENC_f^2} \tag{3.1.48}$$

in which ENC_w represent the white thermal noise, ENC_i the sensor leakage current noise and ENC_f the flicker noise component. Depending on the position of the device, they can be represented with voltage or current generators. Noise sources can therefore be divided into two categories [29]:

- **Series noise**: it is constituted by the voltage noise sources which are connected in series with the amplifier input;
- **Parallel noise**: it is composed of the current noise sources which are connected in parallel with the amplifier input.

As an example, a typical example of parallel noise source is represented by the sensor leakage current, which is modeled as a current source in parallel to the input.
A detailed discussion about the optimization of an amplifier with respect to noise is provided in chapter 4.

3.2 Types of front-end architectures

Depending on the application and the available CMOS technologies, front-end designs have evolved with time. Silicon sensor usage in High Energy Physics experiments has in fact started during the 1970s at CERN and at FNAL, for studies regarding the charm quark. They needed detectors capable of reaching a spatial resolution better than 10 μm, good particle separation and rate capability around 1 MHz [50]. These extreme requirements were not reachable through the traditional detector technologies available at that time, like gaseous detectors. As a consequence, the introduction of silicon detectors, which were able to meet the specifications, was crucial for these studies. At the beginning, silicon strip sensors were developed, while silicon pixel sensors came later. The insertion of these detectors inside collider experiments required the miniaturization of the readout electronics. This purpose has been realized thanks to the development of custom VLSI electronics directly coupled to the sensor. Therefore, starting from the late 1970s a number of techniques for the analog readout of silicon sensors have been developed. In this section the most used methods for analog readout are described, referring to some designs implemented for high energy physics experiments.

3.2.1 Sample-and-hold technique

The sample-and-hold technique consists in the sampling of the amplifier output when a signal is detected. It is then stored into analog memories until it is read out. An example of this approach is represented by the Microplex chip. It is one of the first chips designed for silicon detectors and has been used for the readout of the Silicon Strip Vertex Detector (SSVD) of the Mark II experiment developed at the Stanford Linear Accelerator Center (SLAC). This detector has allowed an improvement in position resolution (around 5 μm per detector layer) leading, among other discoveries, to one of the first measurements of the B lifetime. The SSVD [51], presented in figure 3.11, is composed of 3 layers of silicon detector modules, with 12 modules per layer. Each module features 512 strips and is wire-bonded to four Microplex chips, two at each end. In this way one chip reads out 128 alternate strips [52]. The layout of the chip bonded to the detector is shown in figure 3.12.

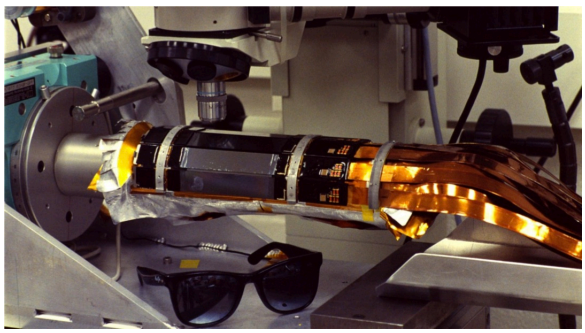

Figure 3.11: Overview of the Mark II vertex detector [50]

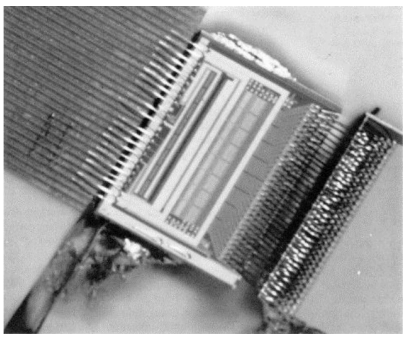

Figure 3.12: Layout of one Microplex chip [53]

The Microplex first version has been developed in 1984 using a 5 μm NMOS technology [54] featuring and active area of $4.4 \times 6.4\ mm^2$. The schematic of a single channel is illustrated in figure 3.13. Two inputs are provided: one from the detector and the other from a calibration line. The amplifier uses a switched capacitor technique. It is designed to have an open-loop gain A equal to 500, a delay time around 10 ns and a rise time of 20 ns. In this case the sample-and-hold works as follows: in the off state, the reset and store switches are closed and the output switches are open [54]. The purpose of the reset switch is that, when it is closed, it

provides an adequate DC bias point for the preamplifier and resets the feedback capacitor. An instant before the data collection begins, the reset is opened. As a consequence, the charge q_{in} coming from the detector is stored on C_{int}. An output voltage equal to q_{in}/C_{int} is produced and saved on C_{store}. In addition, since the effective amplifier input capacitance, which is $(A+1)C_{int}$, is much larger than the typical strip capacitance most of the generated charge will be collected. When data collection ends, the store switch is opened and the reset one is closed. During readout, the switches controlled by the shift register are closed sequentially, one channel at a time, as a low-voltage level makes its way along the read-bit path of the two phase shift register [54]. Furthermore, the voltage of C_{store} appears on the gate of a transistor used to control the output current. The other output line is inserted in order to carry similar switching transients which can be subtracted later using a differential amplifier. Lastly, a DC level, V_{ref}, is provided for pedestal subtraction.

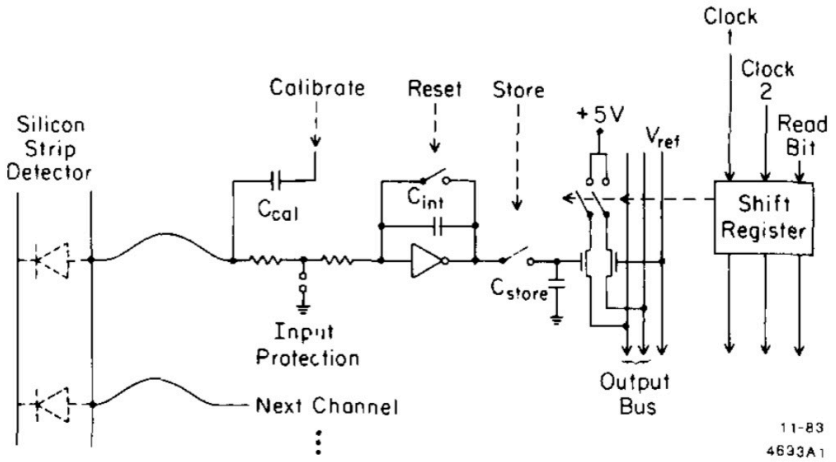

Figure 3.13: Schematic of a Microplex channel [54]

Shortly afterwards, a second version of the microplex chip has been submitted. It has introduced an alternative sample-and-hold method, the correlated double sampling, shown in figure 3.14 [53]. It works as follows. Also in this case, much before data taking the reset and store switches are closed and the output switches are open. Between 0.5 and 1 μs before data collection the reset switch is opened. After the voltages are settled, C_{store1} is opened. As a result, the amplifier output without signal is stored. Then, when the signal arrives, it produces a q_{in}/C_{int} signal at the output of the preamplifier which is stored on C_{store2}. At the end of data collection, also the C_{store2} switch is opened. Then, during readout, the voltages of C_{store1} and C_{store2} appear at the gates of transistors which control the currents flowing into the two output buses. When these currents enter the input of a differential amplifier, everything that produced equal charges on the storage capacitors, such as switching transients, noise from the reset transistor (an important effect), 1/f noise from the input transistor, leakage from the storage capacitors, and channel-to-channel variations in the quiescent output voltage of the amplifiers, are subtracted out. Switching transients from the shift register are also removed. In other words, the main advantage of the correlated double sampling technique is that it allows to get rid of effects which are common to both signal and baseline, improving the quality of the saved data.

The main challenges of this design were not represented by the technology, which was already standard in the 1980s, but by specific aspects of the application. Firstly, the density of connections was 4 times higher than the standard value. In addition, the requirement of an ENC lower than 2000 electrons rms led to explicit constraints on the preamplifier design. Furthermore, a compensation circuit had to be added to prevent oscillations.

Extensive test campaigns have been performed in order to verify the performance of the Mi-

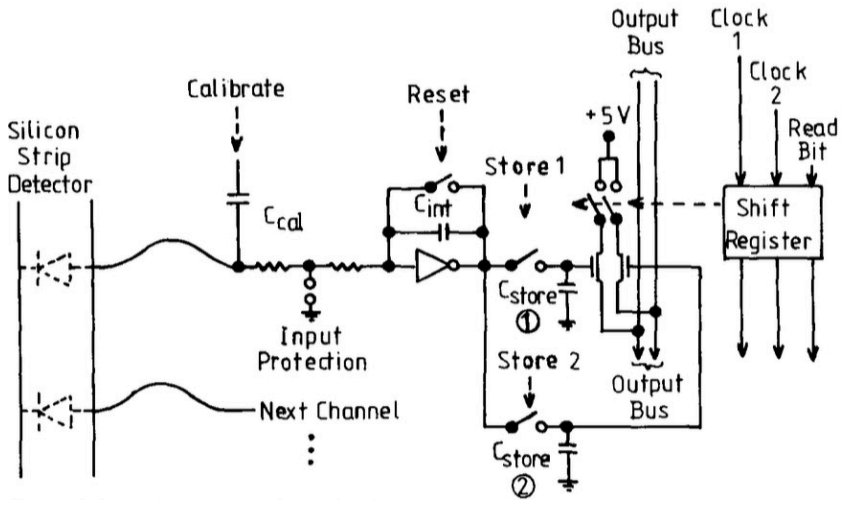

Figure 3.14: Schematic of a Microplex channel featuring correlated double sampling [53]

croplex chip with test beams [55]. In addition, the radiation hardness of the design has been verified. In fact, even if the expected levels of synchrotron radiation at the SLAC linear collider were below 1 krad, it was still a significant requirement at that time [56]. The chips have been irradiated with a ^{60}Co source with an average strength of around 500 rad/hr. The measurements show that the radiation damage has been more severe with the chip under bias, aspect which is commonly expected for CMOS integrated circuits. In fact, with no power applied, the radiation tolerance exceeds the 1 Mrad level, even if one chip failed after 74 krad. Instead, with power applied, the failure occurs in a range between 2.5 and 32 krad. Anyway, it is well beyond the requirement for the MarkII experiment. In addition, an annealing at low temperature has not been able to fully restore the chip performance. Later, a new version with the same functionalities featuring a CMOS process instead of an only NMOS one has been designed [57].

Another ASIC developed during the 1980s and based on this principle is the AMPlifying multipLEXing (AMPLEX) chip [58], designed for the silicon inner detector of the UA2 (Underground Area 2) experiment, located at the CERN Super Proton Synchrotron (SPS) $p\overline{p}$ collider and which has participated in the discovery of the W and Z bosons in 1983. The silicon detector has been included in 1987 to make full use of the enhanced performance of the machine. The detector consists of 192 silicon counters arranged in 12 rows of 16 crystals each, in order to match the granularity and the solid angle coverage, as illustrated in figure 3.15 [59]. It has been inserted to provide a powerful and yet simple tracking device capable of providing a fast reconstruction of the event vertex.
The AMPLEX chip was designed in 3 μm n-well CMOS technology. The main specifications for the design were [58]:

- Limitation of the power consumption to 1 mW per channel in order not to make the cooling system too difficult;
- A signal processing time between 600 and 800 ns to be compatible with the cycling time of 3.5 μs of the $p\overline{p}$ collider.

Figure 3.16 shows the scheme of the 16-channels chip. The charge amplifier is directly connected to the silicon pad detector. It is followed by a shaper, the track-and-hold circuit and the analog multiplexer. The analog circuit design of the charge amplifier and the shaping amplifier is based on the continuous time filtering technique. The preamplifier uses a resistive feedback element R_f to stabilize its DC operating point [58] in place of the reset switch present in the Microplex

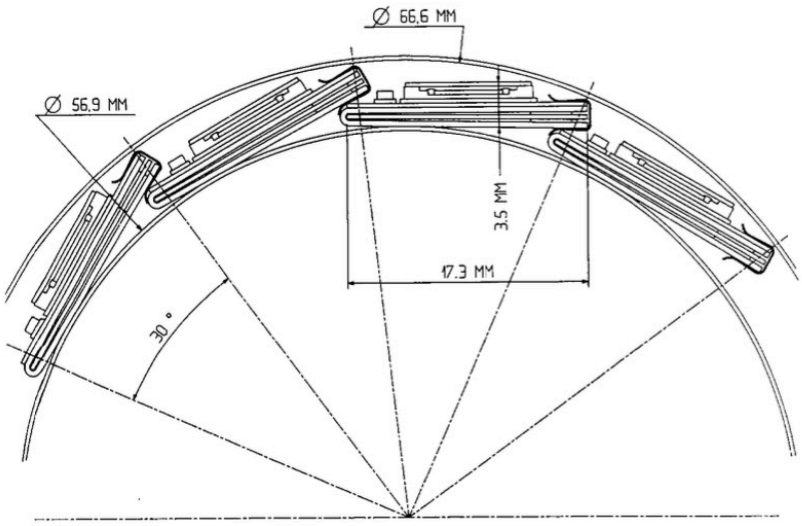

Figure 3.15: Geometrical layout of the inner silicon detector of the UA2 experiment [59]

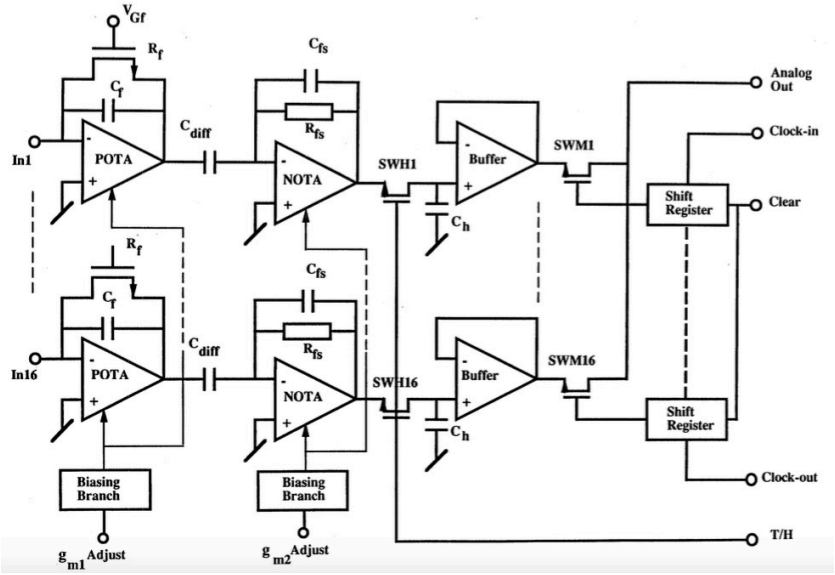

Figure 3.16: Overall schematic of the 16-channel AMPLEX chip [58]

case. Then a semi-gaussian shaping is implemented to optimize noise figures. The following step uses the track-and-hold technique for the storage of the pulse height of the amplified and shaped signal. Lastly, an analog multiplexer consisting of 16 switches is connected to the analog-out bus line.
It is important to underline that sample-and-hold circuits have been quickly replaced in these applications by continuous time architectures, which offered better noise figures. The former approach is in fact quite affected by the charge injections and additional noise caused during the opening and closing phases of the switches. This effect was enhanced by the large size of the transistors available at the time. This trend is already visible in the AMPLEX design, given that the preamplifier is based on continuous time filtering instead of discrete time reset. Nowadays, the shrinking of the transistors dimensions has led to a significant reduction of the charge injections, making again the switched capacitors approaches competitive in analog front-end designs.

3.2.2 Binary front-ends

In HEP experiments the binary readout has been frequently used. It is in fact a simple system, which requires only little digital intelligence. Therefore, binary readout is adequate for systems in which a small silicon area is available. The preamplifier output signal is fed into a comparator which generates a digital pulse, the hit signal, if it goes above a fixed threshold. Nevertheless, only the hit is stored, without information on the analog amplitude. As a result, each signal which crosses the threshold is stored as a good event. For this reason, it is particularly crucial for binary readout systems to minimize the number of hits caused by noise, which requires also an accurate threshold setting.
An example of binary readout is the ALICE1LHCb chip, designed in 2000 for the pixel detector of the ALICE experiment and the RICH detector of the LHCb experiment at CERN. It has been designed with a CMOS 0.25 μm technology, proven to be radiation tolerant up to 30 Mrad, well beyond the 500 krad required for the application. The radiation hardness of the design has been enhanced by using ELTs for NMOS devices and guard rings. In addition, this technology, the most scaled available at that time, allowed a high density of components. The pixel size was in fact quite small, 50×400 μm^2. A block diagram of the analog part of the readout system is illustrated in figure 3.18 [60]. It is composed of a preamplifier and a shaping stage leading to a peaking time equal to 25 ns. All the blocks are differential, having the detector signal and a clean reference as inputs. This choice has been driven by the goal of reducing the digital switching noise contributions injected through the substrate. The possibility of injecting a pulse at the input using the C_{test} capacitance has been included for test purposes. The size of the step is controlled inside the chip to avoid the usage of external signals which can lead to additional noise components.
The shaper output signal is then fed into the discriminator, which has the only purpose of producing a hit if the analog signal goes above the threshold. The discriminator output provides a fast-OR which is used for testing purposes. In addition, a 3-bit register with the corresponding DAC has been included in every pixel in order to provide a fine adjustement of the pixel threshold [61]. The threshold tuning is extremely important, especially in binary systems. In fact, mismatch effects, which start to play an important role in a technology like the CMOS 0.25 μm, lead to different effective threshold between pixels even if the global threshold is the same for the whole chip. In practice this variations act as an additional noise contribution, which can therefore lead to a significant increase of the fake hit rate. The implementation of a local DAC represents the straightforward choice for continuous time applications, in which the correction is always applied. The basic scheme is presented in figure 3.17 [29]. The range of the DAC is defined by the value of the peak-to-peak offset between all the pixels contained in the matrix.

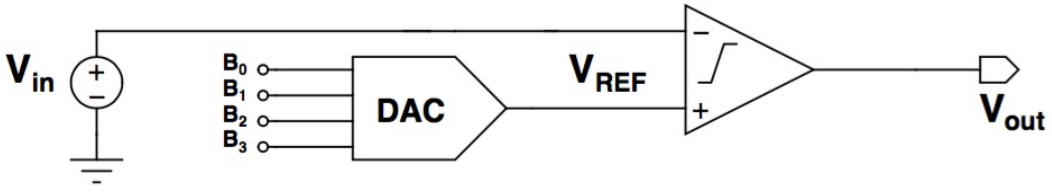

Figure 3.17: Offset compensation via a local tuning DAC [29]

The optimization procedure is defined as follows: at first the threshold of each discriminator is measured; subsequently, the required DAC code for the pixel is calculated and then applied, reducing the threshold dispersion. A local memory is also required to store the bit pattern. The number of DAC bits depends on the application and is usually a trade-off between threshold dispersion minimization and the available area in the pixel.

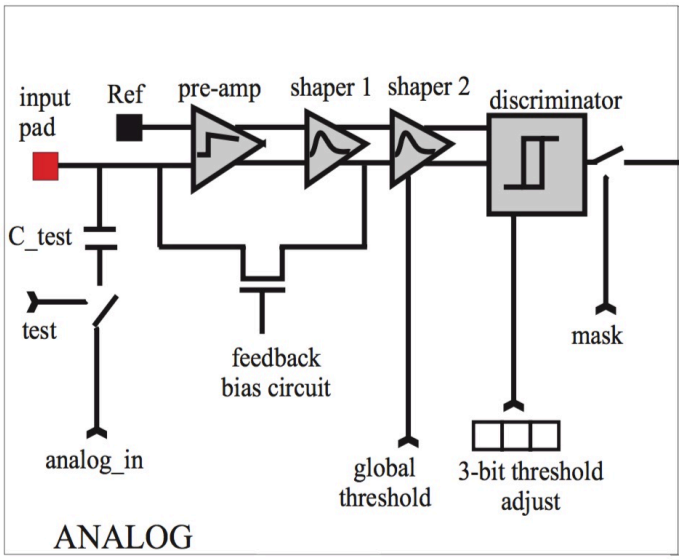

Figure 3.18: Preamplifier-shaper diagram of the ALICE1LHCb chip [60]

3.2.3 Counting and Time-over-Threshold techniques

If the information of interest is the intensity of the radiation, a possibility is to add a counter at the discriminator output. This block can in fact record the number of events in a given period of time. This technique has been implemented into the PX90 chip, designed for X-ray imaging applications. Even if the pixel size is reduced to $100 \times 100\ \mu m^2$ the insertion of a quite complex analog and digital architecture has been possible thanks to the choice of a CMOS 90 nm technology. Figure 3.19 shows the block diagram of a single pixel [62]. In this case the reduction of the noise coming from digital switching activity is performed by implementing a copy of the preamplifier, CSA_REF. The latter is also used to produce the reference voltage for the differential amplifier AMP II. On the other hand, this solution results in a slightly higher electronics noise and power consumption. The AMP II is a fully differential stage which plays a number of roles: it adds a small voltage gain, it features a DAC for the compensation of the threshold dispersion and sets the threshold level for the discriminator. Then the pulses produced by the discriminators are fed into 16-bit counters, which allow to have a very good resolution. Then for the readout the data are latched into output registers that are sequentially addressed. In this way the counters are cleared and a new measurement starts when the data of the previous cycle are transmitted, avoiding dead-time.

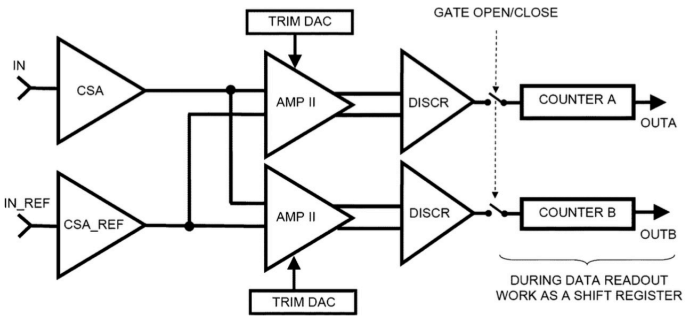

Figure 3.19: Block diagram of a pixel cell unit of the PX90 chip [62] © [2010] IEEE

A binary front-end can be transformed in an amplitude measuring system with few changes. It has to be noted that usually the time spent by the signal over the threshold is proportional to

the amplitude of the input signal. As a result, the duration of the comparator response contains the information about the input charge. An example of this technique is shown in figure 3.20. The amplifier consists of a simple integrator composed of a capacitor C_f connected in feedback loop of a high-gain amplifier [29]. The input pulse generated by the sensor is integrated on C_f, which is in turn discharged by a constant feedback current I_{FEED}. It is then possible to define the time needed to remove the charge:

$$T = \frac{Q_{in}}{I_{FEED}} \tag{3.2.1}$$

The output signal is characterized by a triangular shape and the Time-over-Threshold (ToT) is linearly proportional to the input charge. It is important to underline that the ToT is independent of the value of C_f. In turn, the ratio between I_{FEED} and C_f corresponds to the slope of the triangular signal when it returns to the baseline. Recalling that the peak amplitude is equal to Q_{in}/C_f, if C_f is doubled both the slope and the peak amplitude are halved. As a result, the ToT remains unchanged. In addition, another advantage offered by this method is that even if the amplifier goes in saturation, the linearity of the measurement is not necessarily compromised. It is due to the fact that in saturation the DC gain of the amplifier drops and the virtual ground approximation is not valid anymore. As a consequence, the remaining charge is integrated on the amplifier input node. Nevertheless, it is also removed by the feedback current, leaving the ToT relationship unchanged. It can be altered only in case the residual charge is so large to move the transistors forming the constant current generator out of their working region.

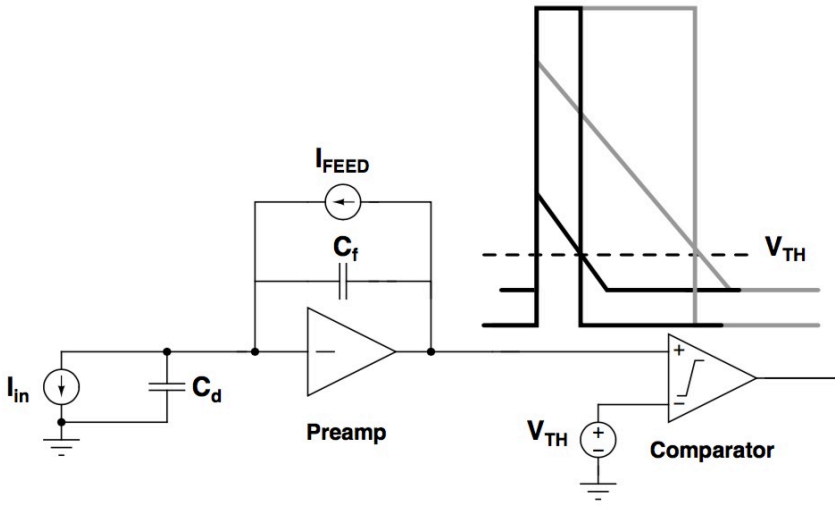

Figure 3.20: Front-end based on a linear Time-over-Threshold measurement [29]

ToT processing is largely used in HEP experiments, given the fact that it allows to have the information related to the input charge. A significant example is the FEI3 chip, designed for the readout of the ATLAS pixel detector. It has been implemented in the CMOS 0.25 μm technology. Also in this application a special care about radiation tolerance was required, given the specification of 50 Mrad of maximum TID. Therefore, ELT devices and guard rings have been used. The pixels, having a size of 50×400 μm, are organized in double columns, as shown in figure 3.21 [63]. The analog chain is similar to the one illustrated in figure 3.20. The preamplifier provides high gain and also the compensation of the sensor leakage current, which becomes significant given the high amount of radiation in this application. It also provides the feedback capacitor discharge required in order to have a correct measurement of the ToT.

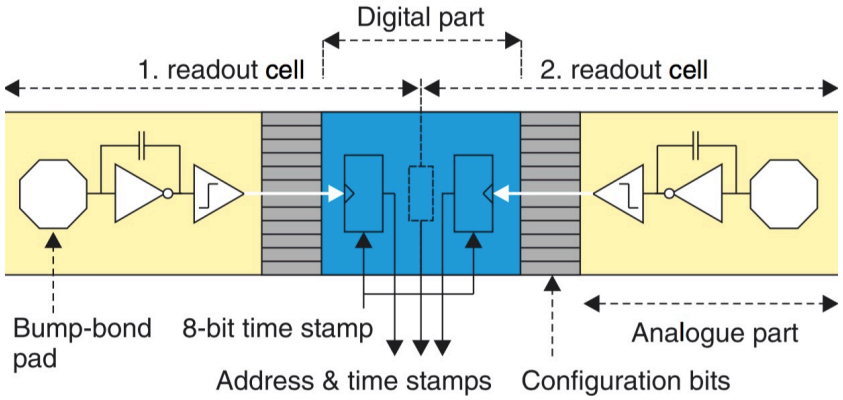

Figure 3.21: FEI3 pixel architecture [63]

Chapter 4

Synchronous front-end design in 65nm CMOS

In this chapter a detailed overview of the analog front-end designed and optimized during the Ph.D. activity is presented. Firstly, the main specifications are listed. Subsequently, each block contained in the analog chain is discussed in-depth, together width the results of the relative CAD simulations.

4.1 The analog front-end scheme

A quick overview of the specifications required for the design of the readout chip for the Phase 2 upgrade of the CMS pixel detector has been already illustrated in chapter 1. This paragraph is focused on the ones that have a direct impact on the analog front-end design:

- **Area**: the pixel size of 50×50 μm^2 puts significant constraints on the size of the devices. In fact, inside RD53 it has been conventionally decided to dedicate 50% of the pixel area to the analog domain and the other half to the digital domain. As a consequence, the choice of the analog building blocks to be included has been also driven by this constraint;

- **Power consumption**: the benchmark on the power consumption of the chip has been fixed at around 0.4 W/cm^2. In this way, the cooling system required to keep the detector temperature low can be maintained relatively simple. Given the small pixel size, it results in a maximum power consumption equal to 10 μW/pixel. Also in this case it has been decided to equally divide it between the analog and digital domains. Therefore, the analog power consumption has to be maintained around 5 μW/pixel;

- **Signal polarity**: even if the final sensor choice is not yet finalized, planar and 3D sensors foreseen for this application are n-type devices. As a result, they both collect electrons, not requiring a bipolar solution for the analog readout;

- **Noise and threshold dispersion**: due to radiation tolerance requirements the sensors that will be used in the HL-LHC experiments are significantly thin and therefore the signal will be small. For this reason, the goal for the RD53 community is to have an analog front-end operating with a minimum threshold equal to 600 electrons. In addition, this value has to be achieved with a 10^{-6} noise hit occupancy, i.e. 0.1 noise hits per bunch crossing in a 10^5 pixels chip. Given that noise is characterized by a normal distribution, this requirement means that the threshold value should be larger than 4.75σ. The overall noise to be considered is the quadrature sum of the ENC and the threshold dispersion, which acts like a supplementary noise term. Therefore, the specification for the front-end

is to obtain:

$$\sqrt{ENC^2 + \sigma_{thr}^2} \leq 126\ e^- \tag{4.1.1}$$

- **Radiation tolerance**: as illustrated in section 2.6, the incident radiation has an important impact on the transistor performances. Since the radiation hardness requirement for the chip is at least 500 Mrad, this aspect needs to be taken into account in the design. Given the high damage suffered by narrow channel PMOS transistors, these devices have been designed with a $W \geq 500nm$. ELT devices have been taken into account in improved versions of the design, illustrated in chapter 6;

- **Hit rate and dead time**: the most internal layer of the tracker is expected to experience a hit rate equal to 3 GHz/cm^2. Therefore the single pixel event rate is 75 kHz. Since in addition a dead time smaller than 1% is required, the signal processing has to be adequately fast in order to be compliant with these specifications.

Keeping in mind these requirements, the analog front-end presented in figure 4.1 has been conceived. It is a synchronous architecture featuring a one-stage preamplifier followed by the discrimination stage. Among the possible choices listed in the previous chapter, this front-end is designed to digitize the ToT information. In fact, this choice gives a good compromise between the requirements of area, power consumption and performance, considering that it can be realized with a quite compact approach. Nevertheless, since a significant part of the signal processing is performed in the analog domain, this design proves to be quite flexible and, if needed, can also be operated in a binary only mode.

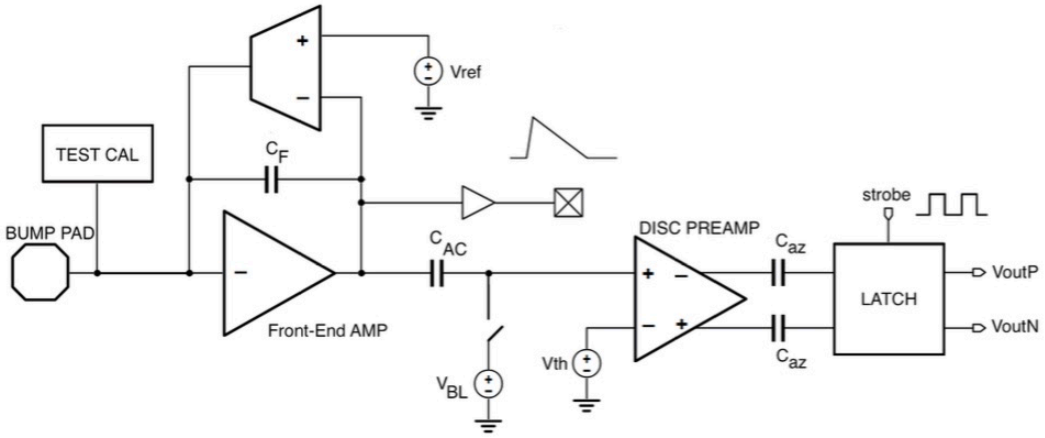

Figure 4.1: Block diagram of the analog front-end

The main blocks are the following:

- **Preamplifier**: its main purpose is to provide the first and main sensor signal amplification. It is designed as a single stage Charge Sensitive Amplifier (CSA) due to the small area and power budget available in the pixel. Its feedback network is realized with the Krummenacher circuit [64]. This choice has been driven by the fact that passive components require too much space in the pixel. Active components, in addition, allow to implement more functionalities. In particular this feedback circuit provides the DC level at the input of the CSA, the sensor leakage current compensation and the constant current discharge of the feedback capacitor. This kind of preamplifier implementation is quite standard in HEP front-end designs. Nevertheless, it is appropriate for this specific application since it is a compact block which provides both amplification and shaping.

- **Calibration circuit**: even if in the final versions of the readout chip the input signal will be provided by the sensor bump-bonded to the chip, for testing purposes it is crucial to include in each pixel a circuit which has the task of injecting a pulse into the preamplifier. This block is indicated by the TEST CAL label in the figure. In addition, this block can be used also in the particle environment for calibration purposes;

- **Synchronous discriminator**: this block is the most innovative section of the front-end design. Its first part consists of a Differential Amplifier (DA) which provides a small additional gain (below 10) and its two inputs are the CSA output signal and the discriminator threshold voltage. The DA offsets caused by mismatch effects are not compensated with the classical threshold trimming DAC. A hardware technique based on the storing of the offset on capacitors is used instead. The usage of a deeply scaled technology like the 65 nm makes in fact again suitable the switched capacitors techniques abandoned during the 1990s. In fact, with a proper sizing of the switches the charge injection effects can be kept under control.
  The two outputs of the DA are then fed into a positive feedback latch which acts as a discrete-time discriminator. In fact, it receives a strobe signal from the digital part which corresponds to the 40 MHz clock. The synchronous scheme has been chosen because HL-LHC is a clocked collider. In particular, the bunches of particles have a time distance of 25 ns. In other words, the clock frequency at HL-LHC is 40 MHz. As a consequence, the collisions in the CMS detector take place every 25 ns. Therefore, given the short distance (few centimeters) between the interaction point and the sensors and the high speed of relativistic particles, a signal is induced on the sensor, and then on the readout chain, only at intervals or multiples of 25 ns. It makes feasible the choice of a discrete-time discriminator. The hit generation is therefore synchronized with the 40 MHz clock, sampling the CSA output around its peaking time. This solution, as a consequence, overcomes the problem of time-walk about time stamp assignment, offering on this topic an advantage with respect to the continuous time architectures.
  In addition, this stage allows a high resolution ToT charge digitization. In fact, by means of an asynchronous logic feedback loop, the latched comparator can be turned in a local Voltage Controlled Oscillator (VCO), whose frequency can be chosen by setting the value of an external bias. As a result, once the CSA signal moves above the threshold, the discriminator starts counting at a much larger frequency, up to 800 MHz from simulations. Therefore, by counting the number of oscillations it is possible to have a precise measurement of the ToT.

A detailed discussion of each of these blocks in provided in the next paragraphs.

4.2 Preamplifier

As explained in chapter 3, the main purpose of the preamplification stage is to amplify the sensor signal by converting it from a current to a voltage. The silicon pixel sensors that will be used at the HL-LHC will be quite thin in order to limit radiation damage. As a result, the expected range of input signals for the readout channel is significantly small, between 1 and 30 ke^-. Therefore, a large gain is required in order to have an adequate signal processing. At the same time, a very low noise configuration is required, in order to be able to detect these small signals while keeping low the fake hit rate.
A schematic of the chosen CSA is illustrated in figure 4.2. It is a transconductance inverting amplifier. The main branch (M1-M4) is composed of a telescopic cascode architecture which allows to obtain a large gain, as required by specifications. In fact, as explained in detail in section 4.2.2, the usage of a cascode for both NMOS and PMOS enhances the large open-loop

Device	Size (W/L) [$\mu m/\mu m$]
M1	8/0.2
M2	2/0.2
M3	2.5/0.25
M4	0.5/2
M5	1.25/0.25
M6	0.5/3
M7	0.5/6
M8	4/0.2

Table 4.1: Transistor sizing of the Charge Sensitive Amplifier

gain. At the same time, a second branch, formed by M5 and M6, is connected at the drain of M1. This technique is known as "current splitting". The idea is in fact to move the most of the current flowing into the input transistor in the second branch. In this way, as illustrated in section 4.2.3, the best compromise between gain and noise is achieved.

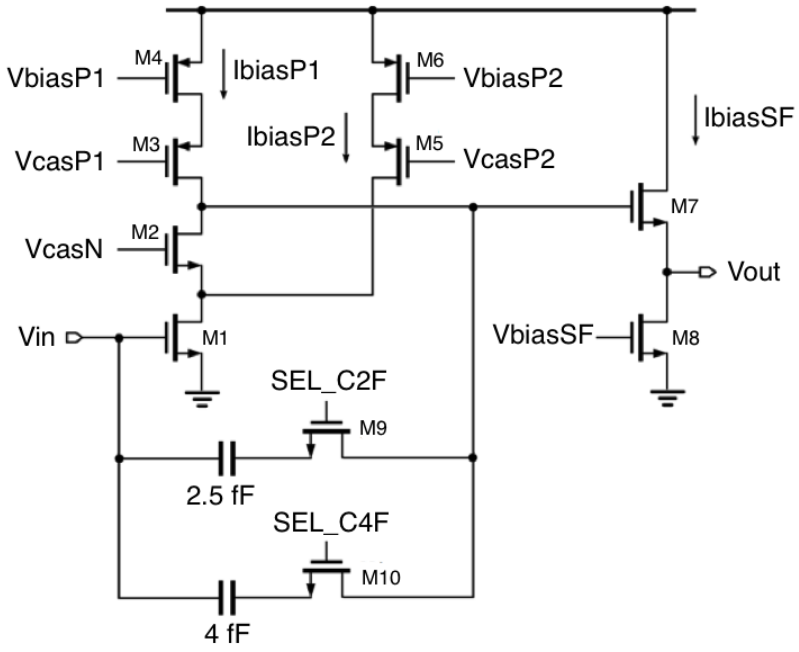

Figure 4.2: Schematic of the Charge Sensitive Amplifier

4.2.1 Transistors and currents sizing

The sizing of the devices is crucial in order to achieve the best compromise between gain, noise and power dissipation. In this case, in addition, also the area is a significant constraint. The dimensions of the transistors are indicated in table 4.1, while the bias current are listed in table 4.2.

The motivations of these choices are exhaustively discussed in the following paragraphs, but they have been anticipated for a better comprehension of some calculations made in the next pages.

Current	Value
IbiasP1	0.5 μA
IbiasP2	1.5 μA
IbiasSF	0.5 μA

Table 4.2: Bias currents of the Charge Sensitive Amplifier

4.2.2 Open-loop gain

Firstly, it is important to make some calculations in order to inspect if this architecture is really able to provide the high gain required by the application. A small signal analysis of this circuit allows to draw fundamental conclusions about the DC gain voltage. The cascode technique is particularly effective for achieving a large gain, since it is characterized by a high output impedance. Both PMOS bias transistors, M4 and M6, have been cascoded. They are therefore characterized by the following small signal resistances [65]:

$$R_{out,cas34} = r_{03} + r_{04} + (g_{m3} + g_{mb3})r_{03}r_{04} \tag{4.2.1}$$

$$R_{out,cas56} = r_{05} + r_{06} + (g_{m5} + g_{mb5})r_{05}r_{06} \tag{4.2.2}$$

The same relationship applies for the NMOS cascode, but it has to be taken into account that in this case the $R_{out,cas56}$ goes in parallel with the output resistance of the input device r_{01}. As a consequence it is possible to write:

$$R_{out,cas12} = (r_{01} \parallel R_{out,cas56}) + r_{02} + (g_{m2} + g_{mb2})(r_{01} \parallel R_{out,cas56})r_{02} \tag{4.2.3}$$

The total impedance at the output of the amplifier is then given by:

$$R_{out} = R_{out,cas12} \parallel R_{out,cas34} \tag{4.2.4}$$

As a result, the small signal DC open loop gain of this stage is:

$$A_0 = g_{m1}(R_{out,cas12} \parallel R_{out,cas34}) \tag{4.2.5}$$

In order to have a high gain it is then also important to obtain a large transconductance of the input transistor. Nevertheless, the telescopic cascode stage, which features four transistors in series, makes M1 working in the weak inversion region. Recalling that in this case the transconductance is given by:

$$g_m = \frac{I_{DS}}{n\Phi_T} \tag{4.2.6}$$

the only way to increase g_m comes by changing the drain current. At the same time, a large current in this stage risks to exceed the power consumption specifications, requiring a trade-off between these two requirements. These aspects motivate the choice of the bias currents. In fact with the values specified in table 4.2 the input transistor current $I_{DS,M1}$ is equal to 2 μA, which is a significant part of the available pixel current budget, around 40%. It leads to a $g_{m1} \simeq 40$ μS which is already a quite remarkable value. Figure 4.3 shows the variation of the resulting open-loop gain with the frequency. The nominal value is compliant with the expectations: it corresponds to 60 dB, which is a factor 1000 in the voltage domain. It remains stable up to around 2 MHz, then it starts to decrease due to the limited bandwidth. It leads to the fact that the CSA is characterized by a non-zero signal rise time. As explained before, it is crucial to assign the event to the correct particle bunch crossing. As a consequence, the CSA peaking time should be less than 25 ns. In order to achieve this result, some constraints on the CSA main parameters have to be set.

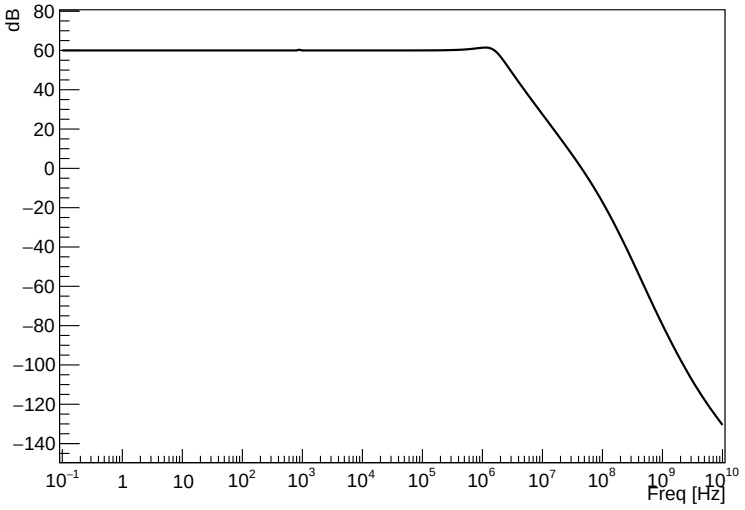

Figure 4.3: Open-loop gain versus frequency

It is then necessary to take into account the feedback capacitance contribution. Figure 4.4 shows the CSA small signal model. The analysis of the small signal circuit has to be started from the equations at the input and output nodes in order to find the transfer function V_{out}/I_{in}:

$$I_{in}(s) + s(C_{det} + C_G)V_{in}(s) + sC_f(V_{in}(s) - V_{out}(s)) = 0 \tag{4.2.7}$$

$$g_m V_{in}(s) + \frac{V_{out}(s)}{R_0} + sC_0 V_{out}(s) + sC_f(V_{out}(s) - V_{in}(s)) = 0 \tag{4.2.8}$$

in which g_m is the transconductance of the input transistor. Solving these equations and remembering that $A_0 = g_m R_0 >> 1$, the result can be simplified in the following way:

$$\frac{V_{out}(s)}{I_{in}(s)} \simeq \frac{1}{sC_f}\left(\frac{1 - sC_f/g_m}{1 + s\tau_r}\right) \tag{4.2.9}$$

It is then composed by one zero and two poles. τ_r is the time constant associated to the second pole and is given by:

$$\tau_r = \frac{1}{g_m}\left[(C_{det} + C_G + C_0) + \frac{(C_{det} + C_G)C_0}{C_f}\right] \tag{4.2.10}$$

Some considerations can be made. The zero is placed at a frequency equal to g_m/C_f due to the direct coupling between input and output through the feedback capacitance. In this design, however, $g_m \simeq 50\ \mu S$ and $C_f \simeq 10\ fF$. As a result, this frequency is in the order of the GHz and the zero can be neglected, leading to a further semplification of the transfer function:

$$\frac{V_{out}(s)}{I_{in}(s)} \simeq \frac{1}{sC_f}\left(\frac{1}{1 + s\tau_r}\right) \tag{4.2.11}$$

By taking the inverse Laplace transform and considering the response to a Dirac delta-like input, this expression can be moved in the time domain:

$$V_{out}(t) = -\frac{Q_{in}}{C_f}\left(1 - e^{-t/\tau_r}\right)u(t) \tag{4.2.12}$$

As a result, at a first approximation the output voltage corresponds to the first-order low-pass filter response. It is then crucial to draw some considerations about the rise time t_r of the

output signal. Conventionally, it is defined at the time needed by the signal to move from the 10% to the 90% of its value. It can be then proven that $t_r \simeq 2.2\ \tau_r$. Looking at equation 4.2.10 a very important contribution is given by g_m. It shows that a large transconductance is also very effective in reducing the rise time, but again a trade-off with power consumption has to be found. In addition, an increase of the feedback capacitance is helpful but diminishes the gain, as shown in figure 4.5.

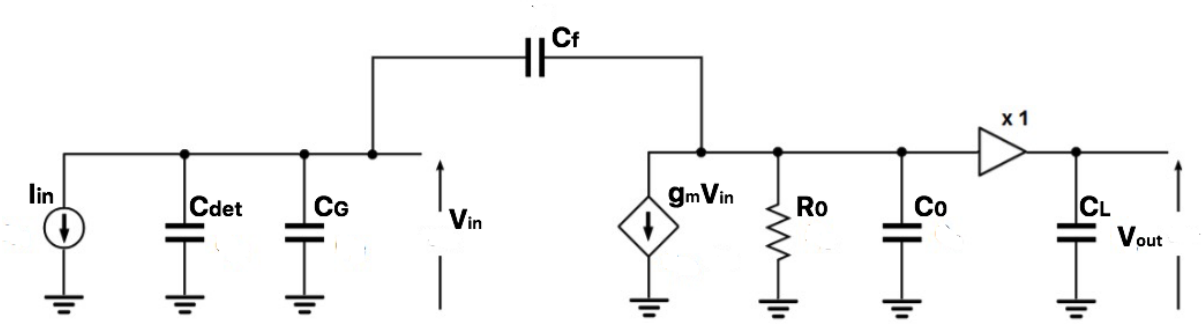

Figure 4.4: Small signal model of the Charge Sensitive Amplifier

Considering the following parameters, which are the default values on this design, the rise time can be estimated:

- $g_m = 42\ \mu S$;
- $C_G = 10\ fF$;
- $C_0 = 4\ fF$;
- $C_{det} = 50\ fF$;
- $C_f = 4\ fF$.

$$t_r = 2.2\tau_r = 2.2\left[\frac{1}{42\ 10^{-6}S}\left(64\ 10^{-15}F + \frac{(60\ 10^{-15}F)4\ 10^{-15}F}{4\ 10^{-15}F}\right)\right] =$$
$$= 2.2\left[\frac{1}{42\ 10^{-6}S}(124\ 10^{-15}F)\right] \simeq 6.5\ ns$$

This value is therefore fully compatible with the requirements, since it is well below 25 ns.

In summary, the telescopic cascode amplifier provides the requested high gain and the transistor and current sizing allows to achieve a rise time well below 25 ns. In addition, since the closed-loop gain is inversely proportional to the feedback capacitance, two capacitors have been included in the design in order to study the behavior of the chain with high gain or small gain.

Source follower

This stage, acting as an output buffer, has been added for a number of reasons. Equation 4.2.10 shows in fact that the rise time of the CSA is strongly dependent on the output capacitance of the stage. In addition, since the CSA output impedance is high, even small variations in the value of the output capacitance can result in significant fluctuations of the CSA bandwidth. Since the output of the CSA has also to drive the feedback network it is then necessary to decouple the cascode output from the feedback network with a buffer, so that the high impedance node is protected.

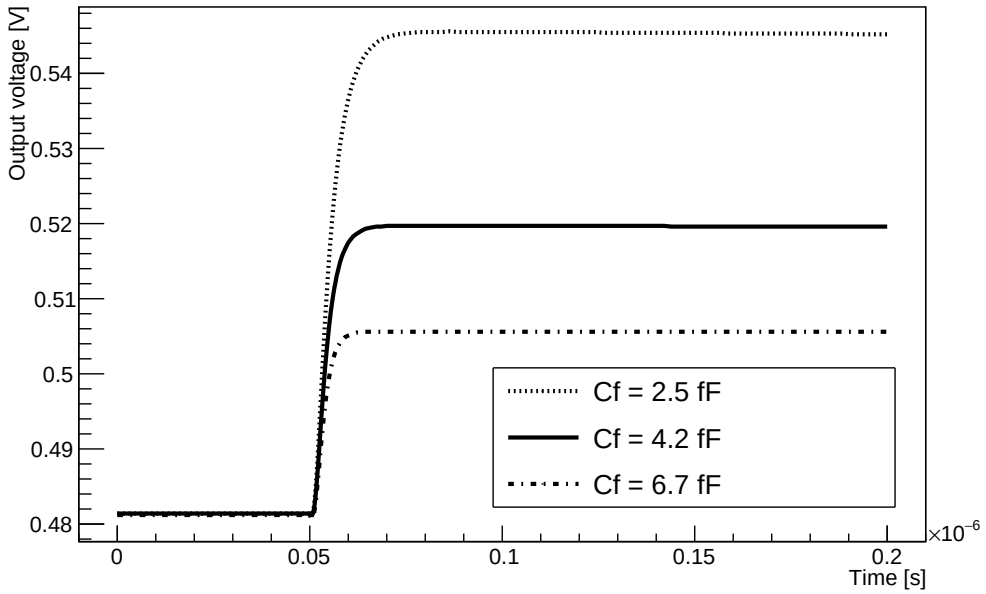

Figure 4.5: Output signal as a function of time for different values of C_f

As a consequence, it is also important to properly size the source follower, since it can have an impact on the CSA performance. As figure 4.6 shows, the rise time of the CSA is dependent on the width of the M7 transistor. Therefore, M7 has been sized with a $W = 4\ \mu m$ since it is a good compromise between rise time optimization and mismatch effects which dominates with small-sized devices.
Figure 4.7 shows instead the dependence of the rise time on the source follower bias current. It is subject to significant variations for small current values, in the order of hundreds of nA, and then it tends to settle to the value imposed by the cascode stage. Keeping in mind the strict power consumption requirements, $I_{sf} = 500\ nA$ has been chosen. In fact, this value of the current results in a very limited increase of the rise time, in the order of 5%, while taking only around 10% of the available power consumption for the analog domain in the pixel.

4.2.3 Noise optimization of a CSA

Each of the transistors inserted in the schematic has an influence on the noise figure. Nevertheless, only some of them produce significant contributions. The main noise source is represented by the input transistor. Therefore its sizing has been defined also taking into account this crucial parameter.

Input transistor sizing

For the optimization of the input transistor sizing it is necessary to analyze the expression of the different noise contributions. Recalling the expressions defined in Chapter 3, the input-referred flicker noise expression for this device is given by:

$$ENC_f^2 = \frac{K_f}{C_{ox}WL} C_{input}^2 N_f \tag{4.2.13}$$

in which N_f is the shaper noise index for flicker noise, and represents a constant quantity which expresses the variation of noise due to the shaping. The input capacitance is given by the sum

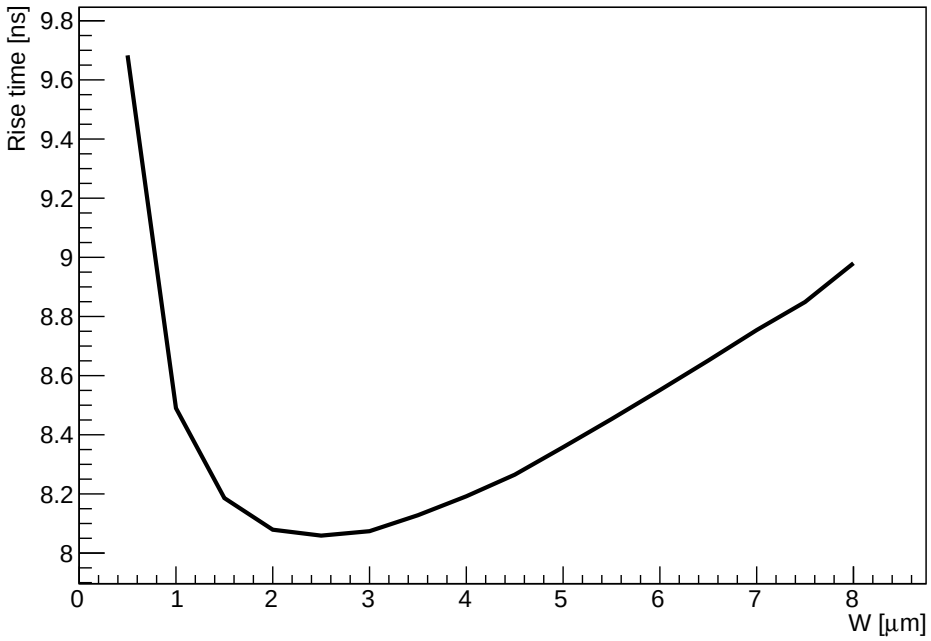

Figure 4.6: Preamplifier rise time as a function of the M7 width

of the detector capacitance and the input transistor gate capacitance. Since the latter is equal to $C_{ox}WL$, the previous relationship can be reshaped as follows:

$$ENC_f^2 = K_f \frac{(C_{det} + C_G)^2}{C_G} N_f \tag{4.2.14}$$

As a consequence, it is now possible to find the condition which minimizes the flicker noise value by differentiating this expression with respect to the gate capacitance, which is controlled by the device size:

$$\frac{dENC_f^2}{dC_G} = N_f K_f \left(\frac{2(C_{det} + C_G)C_G - (C_{det} + C_G)^2}{C_G^2} \right) = 0 \tag{4.2.15}$$

The condition which solves this equation is therefore:

$$C_G = C_{det} \tag{4.2.16}$$

As a result, a gate capacitance equal to the detector one minimizes flicker noise contributions. Regarding thermal noise, the spectral density is instead:

$$v_{nw}^2 = \frac{4k_B T n \gamma \alpha_w}{g_m} \tag{4.2.17}$$

in which n is the body factor, γ is the inversion factor introduced in chapter 2, α_w the excess noise factor and g_m is the input transistor transconductance. The input-referred ENC due to thermal noise is described by:

$$ENC_w^2 = 4k_B T n \gamma \alpha_w \frac{(C_{det} + C_G)^2}{g_m(C_G)} \frac{N_w}{T_P} \tag{4.2.18}$$

in which N_w is the shaper noise index for thermal noise and T_P is the peaking time. The g_m dependence on C_G is written in order to underline that the gate capacitance enters this

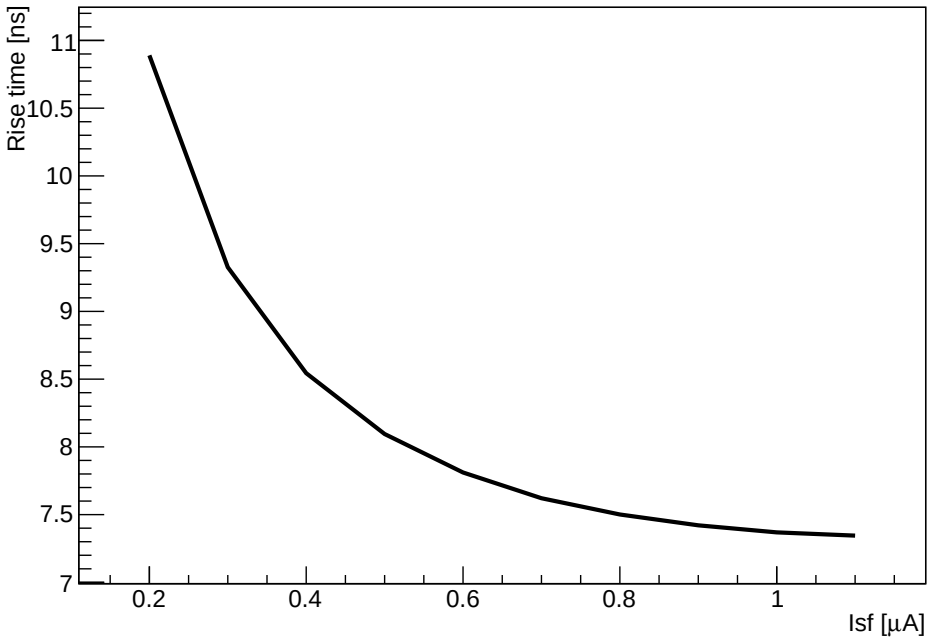

Figure 4.7: Preamplifier rise time as a function of the source follower bias current

expression in two opposite directions. Therefore, also in this case a trade-off has to be found, also in this case by differentiating with respect to C_G:

$$\frac{dENC_w^2}{dC_G} = \frac{4k_BTn\gamma\alpha_w N_w}{T_P}\left[\frac{2(C_{det}+C_G)g_m - (C_{det}-C_G)^2\frac{dg_m}{dC_G}}{g_m(C_G)^2}\right] = 0 \tag{4.2.19}$$

It can be simplified obtaining the following expression:

$$2g_m = (C_{det}+C_G)\frac{dg_m}{dC_G} \tag{4.2.20}$$

This is a general formula. To understand more, it is necessary to find the expression of g_m as a function of C_G. Considering as an example a transistor in strong inversion, the transconductance is given by:

$$g_m = \sqrt{2\mu C_{ox}\frac{W}{L}I_{DS}} \tag{4.2.21}$$

Remembering that C_G can be approximated with $C_{ox}WL$. By multiplying and dividing then equation 4.2.21 by L it is possible to write:

$$g_m = \sqrt{2\mu C_G\frac{I_{DS}}{L^2}} \tag{4.2.22}$$

Therefore differentiating g_m with respect to C_G the following expression is obtained:

$$\frac{dg_m}{dC_G} = \frac{\mu\frac{I_{DS}}{L^2}}{\sqrt{2\mu C_G\frac{I_{DS}}{L^2}}} \tag{4.2.23}$$

This result can be inserted in equation 4.2.20:

$$2\sqrt{2\mu C_G\frac{I_{DS}}{L^2}} = (C_{det}+C_G)\frac{\mu\frac{I_{DS}}{L^2}}{\sqrt{2\mu C_G\frac{I_{DS}}{L^2}}} \tag{4.2.24}$$

$$4\mu C_G \frac{I_{DS}}{L^2} = (C_{det} + C_G)\mu \frac{I_{DS}}{L^2} \tag{4.2.25}$$

$$4C_G = C_{det} + C_G \tag{4.2.26}$$

As a consequence, the final expression of the gate capacitance which minimizes thermal noise in strong inversion is:

$$C_G = \frac{1}{3} C_{det} \tag{4.2.27}$$

In case of an input transistor biased in strong inversion the strategy can therefore be the following. Firstly, it is important to know the maximum power consumption and the peaking time specifications. Then the maximum current compatible with this requirement together with the minimum L allowed by the technology are chosen, since they increase the value of g_m, as shown in equation 4.2.21. At this point the gate capacitance value is fixed according to 4.2.27 by changing the W of the device. If then the flicker noise is higher than the thermal one, W can be further increased in order to find the best compromise [29].
Nevertheless, it should be taken into account that a large increase of the W with a fixed current leads to an important decrease of the overdrive voltage $V_{GS} - V_{TH}$. As a consequence, at some point the device changes its region of operation, moving towards moderate and weak inversion. In the latter case, the transconductance is:

$$g_m = \frac{I_{DS}}{n\phi_T} \tag{4.2.28}$$

Therefore, it is independent on the gate capacitance. As a result, a too large increase in W does not lead to a large improvement of the noise figure. Since in deep submicron technologies like 65 nm the transistors tend to work in weak or moderate inversion, as it is the case of M1 in the CSA, the previous method of optimization proves to be inadequate for this application. A more complicated procedure is required. In order to find the ENC expression in the weak inversion case, it is useful to recall the concept of inversion coefficient presented in chapter 2, since it allows to express the relevant quantities continuously across the different levels of inversion. It allows to rewrite the transconductance as follows:

$$g_m = \frac{I_{DS}}{n\phi_T} \frac{1}{\sqrt{I_C + 0.5\sqrt{I_C} + 1}} \tag{4.2.29}$$

The total gate capacitance can be expressed as follows:

$$C_G = C(x) C_{ox} W L \tag{4.2.30}$$

in which the coefficient C(x) is given by:

$$C(x) = \frac{n - (1 + x)/3}{n} \tag{4.2.31}$$

x is in turn dependent on the inversion coefficient:

$$x = \frac{(\sqrt{I_C + 0.25} + 0.5) + 1}{(\sqrt{I_C + 0.25} + 0.5)^2} \tag{4.2.32}$$

In this way, the quantities included in the ENC expression are valid in all regions of operation. In addition, for the overall noise it is necessary to take into account also the contributions due to the sensor leakage current:

$$ENC_i^2 = 2qI_{leak} N_i T_P \tag{4.2.33}$$

in which N_i is the shaper noise factor for parallel noise.
Furthermore, some aspects which are specific of deep submicron technologies have to be included in order to perform an accurate input transistor sizing. In fact, usually the usage of L_{min} devices is not recommended. Firstly, it leads to a significant reduction of the output conductance, limiting the DC gain of the amplifier. In addition, when the minimum channel length is used the excess noise factor of white noise α_w increases, together with the flicker noise coefficient K_f. Concerning PMOS transistors, a dependence of K_f on the current density and a deviation from the $1/f$ behavior of the flicker noise have been demonstrated. Therefore, this process is better described by the following power spectral density:

$$v_{nf}^2 = \frac{K_f(L, I_C)}{C_{ox}WL}\frac{1}{\alpha_f} \tag{4.2.34}$$

The exponent α_f is dependent on the device:

- $\alpha_f = 0.85$ for PMOS transistors;
- $\alpha_f = 1.1$ for PMOS transistors.

Another aspect to be considered is the gate overlap capacitance. After fabrication, in fact, the drain and source terminals have a small overlap with the gate contact. Its value is around 1 fF per micron of transistor width. As a result, it forms a significant contribution in the overall gate capacitance, which can be rewritten as follows:

$$C_G(I_C, W, L) = C(x)C_{ox}WL + 2C_{ov}W = C_{GW}(I_C, L)W \tag{4.2.35}$$

in which C_{GW} is the gate capacitance per unit of width.
It is then possible to write the final expression of the ENC, based on the fact that

$$ENC^2 = ENC_f^2 + ENC_w^2 + ENC_i^2 \tag{4.2.36}$$

Inserting the different contributions, it becomes:

$$ENC^2 = (C_{det} + C_{GW}(I_C, L)W)^2 \left[\frac{4k_BTn\gamma\alpha_w}{g_m(I_C)}\frac{N_w}{T_P} + \frac{K_f(L, I_C)}{C_{ox}WL}\frac{N_f(\alpha_f)}{T_P^{(1-\alpha_f)}}\right] + 2qI_{leak}N_iT_P \tag{4.2.37}$$

Based on this relationship, the optimization procedure works as follows. It should be started by determining the value of specifications like the input capacitance, the detector leakage current, the event rate, the maximum available power consumption and the target ENC. The peaking time is then determined from the rate requirements and the ENC due to parallel noise is calculated. If the latter gives already a small contribution the peaking time can be considered defined, otherwise it should be reduced.
Let's now apply the main concepts to the design. Figure 4.8 shows that the optimum value of the gate width for the input transistor is larger than 4 μm. It is a quite large value, especially for a 65 nm technology. Usually such a large device is not laid out as a single transistor. One of the main reasons is about noise. In fact, the gate material is characterized by a gate resistance R_G. It therefore produces a thermal noise component that can be modeled by a voltage source equal to $4k_BTR_G$ which is connected in series to the input. If then the transistor is split in multiple devices connected in parallel, R_G is smaller. In addition, the two end-points of each finger are shorted with a metal layer to further reduce the resistance.

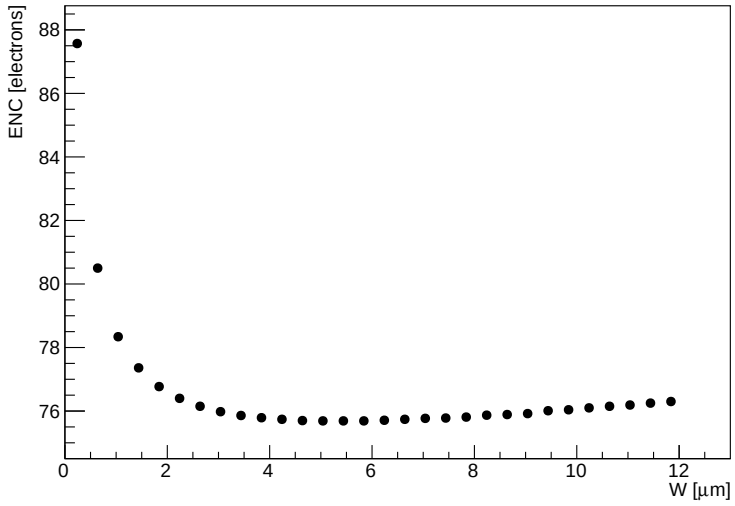

Figure 4.8: Equivalent Noise Charge as a function of the input transistor (M1) gate width

Other devices

Regarding the other transistors, an important contribution is given by the biasing current sources. Let's consider a case in which the splitting branch is not included. M4 would be characterized by a channel thermal noise given by:

$$i_{n4}^2 = 4k_B T \alpha_{w4} \gamma_4 g_{m4} \tag{4.2.38}$$

Also this value can be referred to the input by dividing it by the transconductance of M1:

$$v_{nw4}^2 = 4k_B T \alpha_4 \gamma_4 \frac{g_{m4}}{g_{m1}^2} = 4k_B T \alpha_4 \gamma_4 \frac{1}{g_{m1}} \frac{g_{m4}}{g_{m1}} \tag{4.2.39}$$

As a consequence, the thermal noise contribution rising from the biasing device is practically the same of the input transistor multiplied by the ratio of the transconductances. Then to minimize this contribution g_{m4} should be significantly smaller than the one of g_{m1}. Without splitting branch, M1 and M4 are biased with the same current. As a result, the only suitable method to reduce g_{m4} is to decrease the W/L, pushing the transistor deeply in strong inversion. Nevertheless, this procedure increases the overdrive voltage and therefore the headroom required to keep the transistor in saturation, reducing the one of M3, which sustains the signal. In addition, it can reduce the output conductance of M4. In other words, this procedure has an impact on the output gain of the amplification stage.
The current splitting architecture is therefore advantageous on this side. In fact, the M5-M6 branch is designed in a way that it drives the 75% of the current flowing in the input transistor, thus reducing the current flowing in M4. As a result, it is sufficient to size the latter such that it works at the onset of strong inversion, without big impact on signal processing. In addition, M6 becomes the main noise source among the bias sources. Nevertheless, it can be put in deep strong inversion since this branch is not significantly involved in large signal swings.
Concerning cascode transistors, they bring a very small contribution to the overall noise figure. It is due to the fact that their transconductance is degenerated by the output conductance of the load [29]. As an example, the transconductance of M3 is given by:

$$g_{m3,eq} = \frac{g_{m3}}{1 + g_{m3} r_{04}} \tag{4.2.40}$$

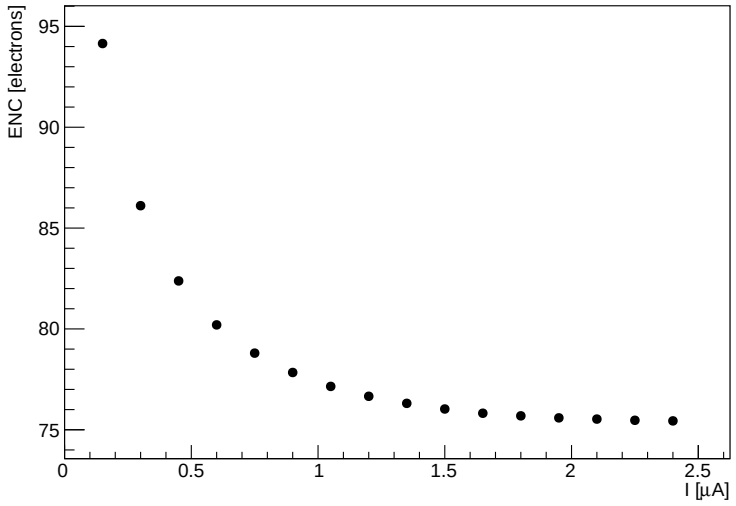

Figure 4.9: Equivalent Noise Charge as a function of the input transistor (M1) drain-source current

4.2.4 Krummenacher feedback

Regarding the feedback network, an active scheme has been chosen. The idea of implementing it with a passive component was in fact not viable for both its functionality and its size, difficult to include in the small pixel area available. Among the possible choices the Krummenacher feedback [64] scheme has been used, because it allows to merge multiple functionalities in the same block:

- It provides the DC level at the input of the CSA;
- It allows the compensation of the sensor leakage current;
- It provides a constant current discharge of the feedback capacitance.

The schematic of this block is presented in figure 4.11. The main part is composed of a differential stage with an auxiliary branch which provides the correct biasing for the M4 transistor. The input differential pair has been designed with NMOS devices in order to be compatible with the negative polarity of the input signal. The drain of M1 is connected to the amplifier input. This part corresponds to a resistor of value $R_f = 1/g_{m1}$ in parallel with the feedback capacitance C_f. The other branch, which contains the M2 transistor, is characterized by a drain current which is integrated on the capacitor C_C. It leads to a voltage which controls the gate of the PMOS M5. This second feedback path behaves as an inductor connected in parallel with C_f. As a consequence the DC component of the sensor leakage current flows into M5 rather than into R_f. The advantage is that M5 can sink a total current larger than $I_{krum}/2$, which is the bias current of the feedback path. As a result, the leakage current can exceed, up to a large limit, the value of $I_{krum}/2$ without compromising the behavior of the preamplifier. This feature is confirmed in figure 4.12. It shows in fact that the preamplifier output experiences very small variations even for values of the leakage current around 100 nA, i.e. well beyond the expected levels at HL-LHC.
At the same time it has to be underlined that I_{krum} and R_f contribute to the parallel noise together with the leakage current. Figure 4.13 shows in fact that the ENC linearity versus the sensor capacitance is still verified, but compared to the case presented in figure 4.10 realized with an ideal feedback the slope increases. Considering a standard sensor capacitance of $50fF$,

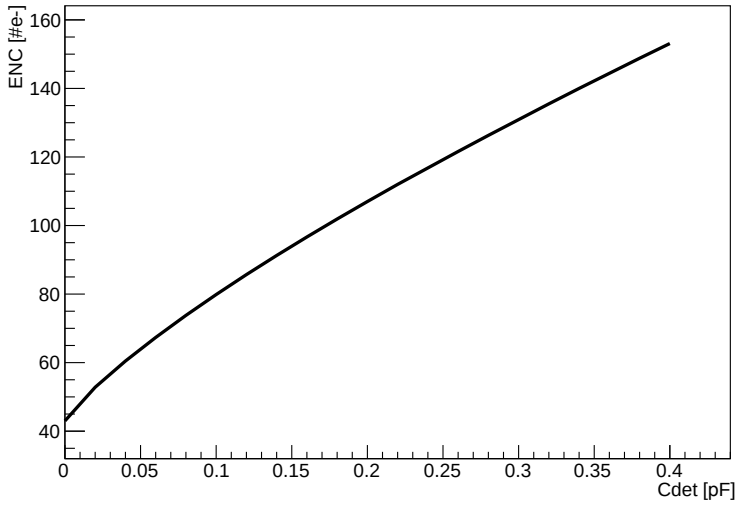

Figure 4.10: Equivalent Noise Charge as a function of the detector capacitance

Device	Size (W/L) [$\mu m/\mu m$]
M1	1.2/1
M2	1.2/1
M3	0.4/8
M4	0.5/3
M5	0.5/3
M6	0.4/8
M7	0.2/3
M8	0.5/3

Table 4.3: Transistor sizing of the Krummenacher feedback network

the additional contributions due to the Krummenacher current is illustrated in figure 4.14. The ENC tends in fact to linearly increase with I_{krum}. At the same time, also the sensor leakage current causes an increase of the ENC. The variations are presented in figure 4.15 for a sensor capacitance equal to 50 fF and a Krummenacher current of 20 nA.
In addition, the correct damping of the feedback loop has to be guaranteed. This purpose is achieved if the transconductance of M5 respects the following condition:

$$\frac{C_C}{g_{m5}} > 2\frac{C_f}{g_{m1}} \tag{4.2.41}$$

g_{m5} is dependent on the sensor leakage current. As a result, this condition has to be respected for the maximum leakage current. The transistor sizing shown in table 4.3 has been therefore driven also by this considerations.
Running a simulation with a sensor leakage current equal to 20 nA, which is the maximum expected value after irradiation, it is possible to find that:

- $C_C = 536\ fF$
- $g_{m5} = 485\ nS$
- $C_f = 2.5\ fF$

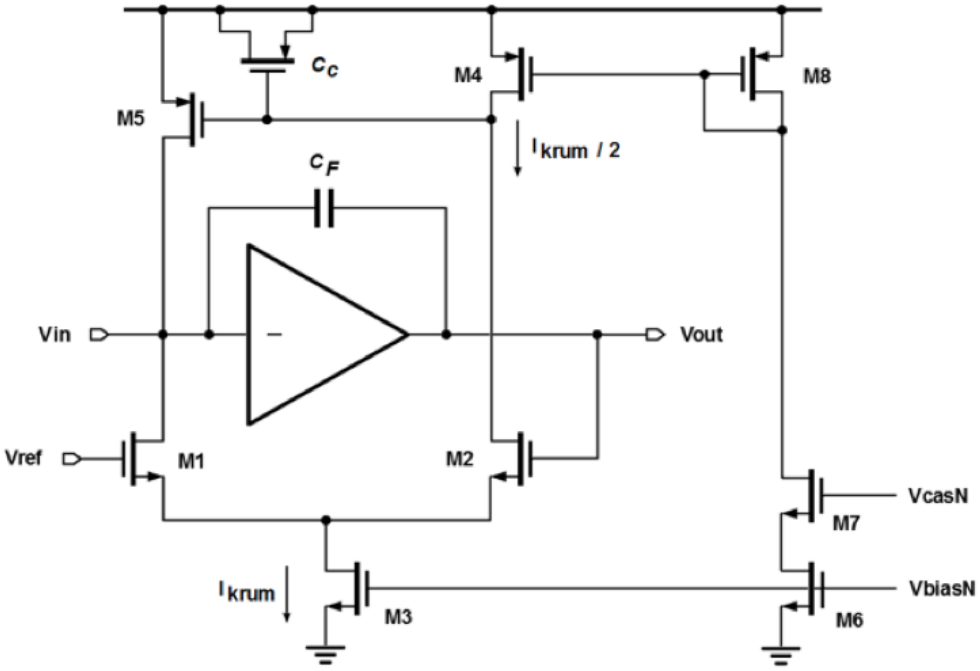

Figure 4.11: Schematic of the Krummenacher feedback adopted

- $g_{m1} = 15\ nS$

These values lead to:

$$\frac{C_C}{g_{m5}} > 2\frac{C_f}{g_{m1}} \rightarrow \frac{536\ fF}{485\ nS} \simeq 10^{-6} > \frac{2.5\ fF}{15\ nS} \simeq 1.6 \times 10^{-7} \tag{4.2.42}$$

Therefore, the requested condition is respected.
As explained before, the value of I_{krum} determines also the constant current values with which the feedback capacitor is discharged. Therefore this parameter influences the Time-over-Threshold value. Since here only the first stage is considered, an "analog" ToT value is computed by superimposing a threshold in the simulation corresponding to the amplitude generated by a 600 e^- input signal. The latter is in fact the reference value for the real threshold. An example of the impact of I_{krum} on the ToT is presented in figure 4.16 for a 10 ke^- signal. This parameter has a quite significant impact, especially for small currents. In fact:

- I_{krum} = 10 nA $\rightarrow$ ToT = 310 ns;
- I_{krum} = 40 nA $\rightarrow$ ToT = 90 ns;

Keeping in mind the requirement of a dead time smaller than 1 %, it is therefore important to find a value of the feedback capacitor constant current discharge compatible with it. On the other hand, also the noise value has to be kept under control. Therefore a good compromise between these aspects is given by a value of feedback current between 20 and 40 nA. Beyond the latter, in fact, the gains in ToT are small, while the ENC linearly increases. In turn, values of the Krummenacher current under 20 nA lead to large increase of the ToT value, which can be an issue at least for the innermost layer which experiences the 3 GHz/cm^2 rate.
In addition, the value of the Krummenacher current has an impact on the preamplifier output signal shape. This aspect is clarified by figure 4.17 which represents the output voltage for a 10 ke^- input charge as a function of time. The fast discharge configuration (40 nA) leads to some

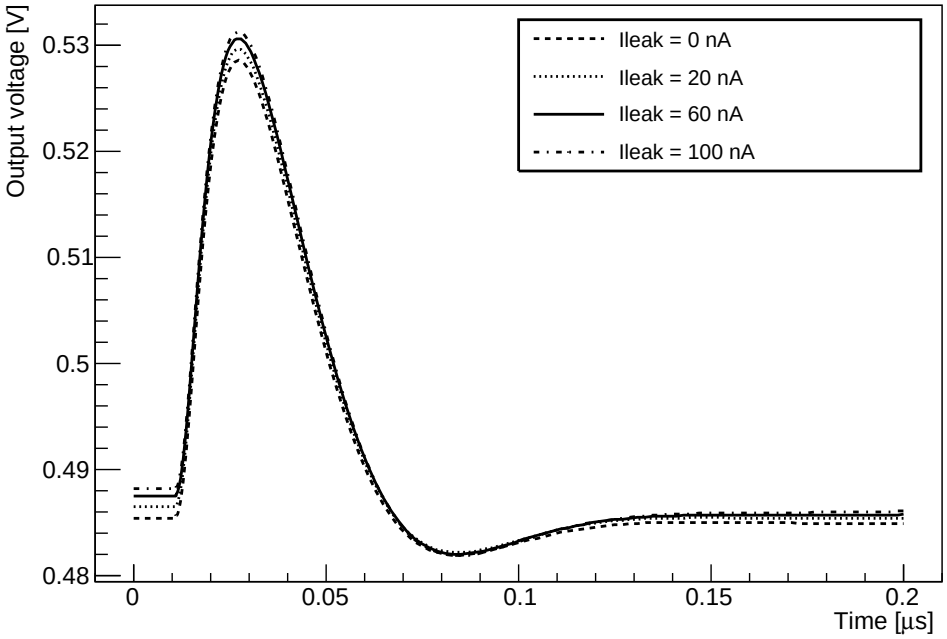

Figure 4.12: CSA output for a 1 ke^- signal with different leakage current values

secondary aspects that have to be taken into account. Firstly, the amplitude is slightly reduced. In fact, the Krummenacher current works also during the rising edge phase and a small part of the charge is removed before the peak is reached. The higher is the feedback current the larger this effect, called "ballistic deficit" takes place. Furthermore, an increase of I_{krum} leads to a larger instability. Therefore at the end of the discharge period the signal tends to move below the baseline ("undershoot") and then slightly above ("overshoot"), signaling the beginning of a ringing effect. The overshoot can especially represent an issue if it is significantly large, since it is superimposed to noise and can lead to an increase of the fake hit rate. WIth a 40 nA current both the ballistic deficit and the undershoot/overshoot effects are under control. Nevertheless, they represent a limit in a further increase of the Krummenacher current, together with noise.

A last aspect to be underlined about the preamplifier is that the high gain provided together with the small maximum voltage of this technology, 1.2 V, leads to a saturation of the output voltage amplitude beyond an input charge of 10 ke$^-$. This feature is illustrated in figure 4.18. Nevertheless, in this design the information which is digitized is not the signal amplitude, but the Time-over-Threshold. Then, as figure 4.19 shows, this architecture allows to have a very good linearity of the ToT in the whole range of interest about the input charge. Only a small deviation is experienced at very low values of Q_{in}.

4.2.5 Calibration circuit

The calibration circuit has the purpose of injecting a test current pulse in the preamplifier. The schematic, shown in figure 4.20, has been mutuated by similar solutions adopted in other HEP applications [66]. The $TestP$ digital signal is common between all pixels, while the CAL_EN one is at the pixel level and enables the pulse injection.
Depending on the values of $TestP$ and CAL_EN one of the two CMOS switches, SW1 or SW2 is enabled. In fact, SW1 is controlled by an AND gate while SW2 by a NAND gate. The NMOS devices into the switches are preceded by a delay cell to match the delay caused by the inverter on the PMOS gate in order to minimize charge injections. The configuration is

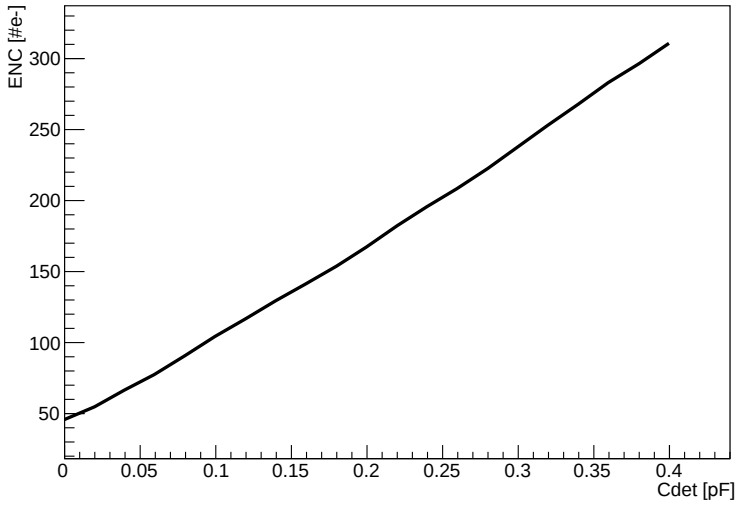

Figure 4.13: ENC versus the sensor capacitance with the Krummenacher feedback inserted

$TestP$	CAL_EN	Switch enabled
0	0	SW2
0	1	SW2
1	0	SW2
1	1	SW1

Table 4.4: Configuration scheme of the calibration circuit

summarized in table 4.4.
Two DC voltage levels, common between all the pixels, are then used, CAL_HI and CAL_LO, with the former larger than the latter in the following way. Given, as a starting condition, that $CAL\_EN = 1$ and $TestP = 0$, CAL_LO is connected to the injection capacitor C_{cal}. When $TestP$ becomes high, at a time t_0, CAL_LO is disconnected and at the same time the capacitor is tied to CAL_HI. As a result, a positive pulse is injected into the preamplifier. Since the latter is an inverting stage, it results in a negative output voltage. On the other hand, when $TestP$ returns low at a time t_1, a negative pulse is sent to the preamplifier. Therefore, this is the signal of the desired polarity. At the same time, the length of the $TestP$ signal has to be properly tuned in order to avoid that the wrong polarity signal is still processed when the right one is injected. The amount of the injected charge is well explained by the following relationship:

$$Q_{cal} = \int_{t_0}^{t_1} i(t)dt = \int_{CAL\_LO}^{CAL\_HI} C_{cal}dV_c = C_{cal}(CAL\_HI - CAL\_LO) \tag{4.2.43}$$

Therefore, since the input charge depends only on the difference between the two voltage levels, it is also possible to fix the value of CAL_LO and change the CAL_HI one. Furthermore, in order to minimize the number of lines coming from the chip periphery, CAL_LO has been tied to ground. As a result, the final expression of the input charge is the following:

$$Q_{cal} = C_{cal}CAL\_HI \tag{4.2.44}$$

Recalling that in the 65 nm technology the maximum voltage is 1.2 V, this value represents also the limit of voltage difference at the injection capacitance input. Therefore, in order to be

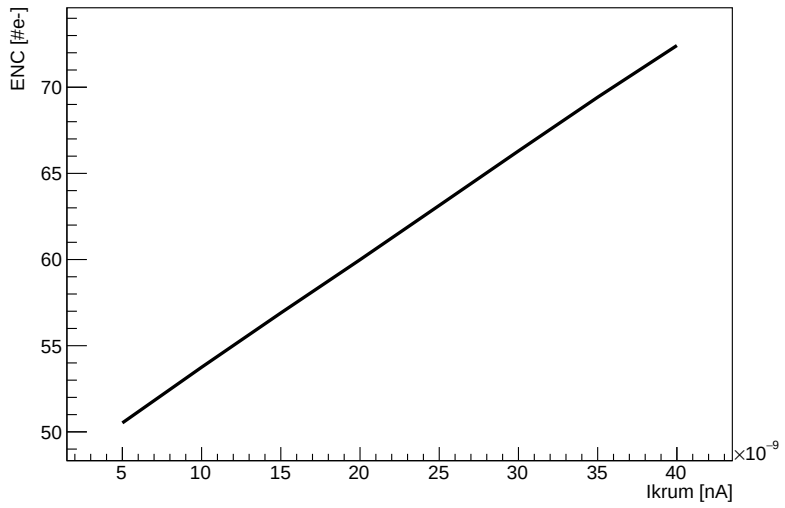

Figure 4.14: ENC versus the Krummenacher current for a $C_{det} = 50fF$

able to exploit all the desired range on input charge (1 - 30 ke$^-$) the size of the C_{cal} should be properly determined. As a consequence, $C_{cal} \simeq 8\ fF$ has been chosen. Therefore:

$$Qin = 1ke^- \rightarrow CAL\_HI = \frac{Q_{in}}{C_{cal}} = \frac{1000\ e^-}{6250\ e^-/fC}\ \frac{1}{8\ fF} = 20\ mV \tag{4.2.45}$$

$$Qin = 30ke^- \rightarrow CAL\_HI = \frac{Q_{in}}{C_{cal}} = \frac{30000\ e^-}{6250\ e^-/fC}\ \frac{1}{8\ fF} = 600\ mV \tag{4.2.46}$$

As a result, this sizing allows to inject all the required charges while keeping a good margin from the 1.2 V.

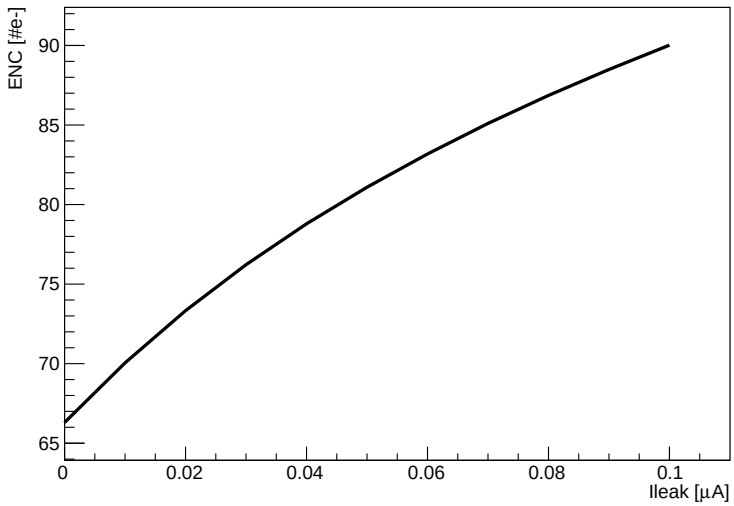

Figure 4.15: ENC versus the leakage current for a $C_{det} = 50fF$

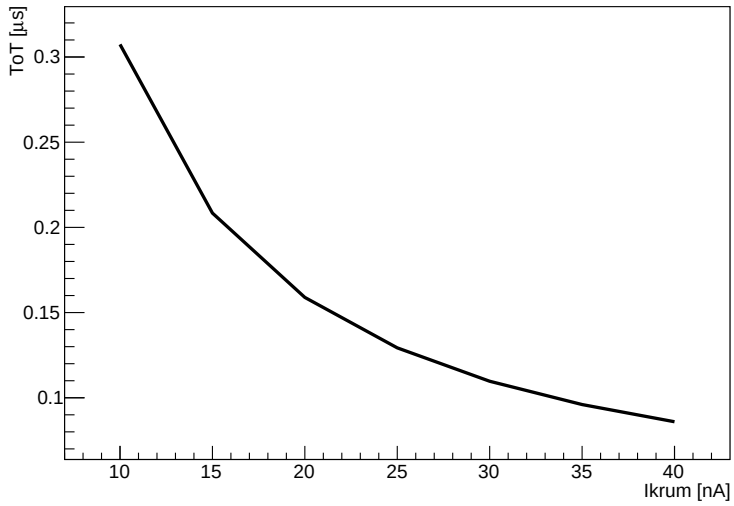

Figure 4.16: ToT as a function of the Krummenacher current for a 10 ke- input signal

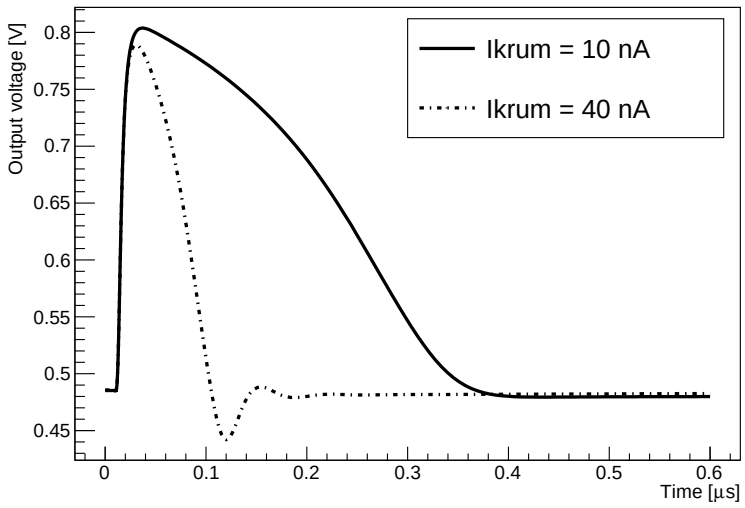

Figure 4.17: Preamplifier output voltage for a 10 ke$^-$ input charge with different values of I_{krum}

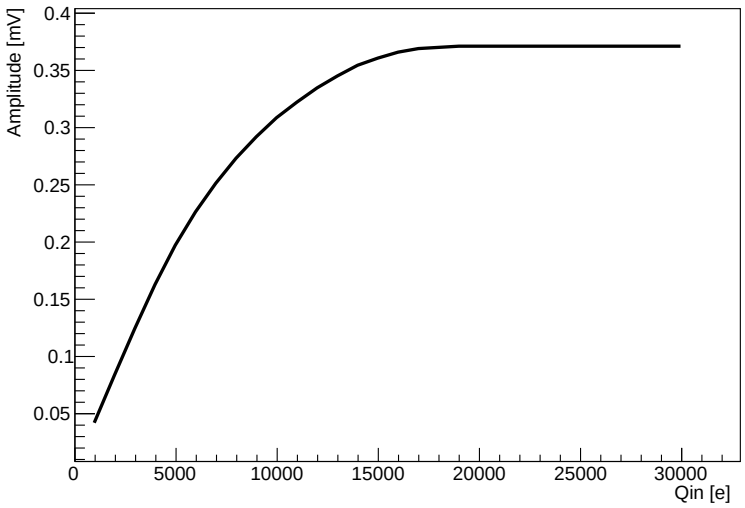

Figure 4.18: Preamplifier output signal amplitude as a function of the input charge

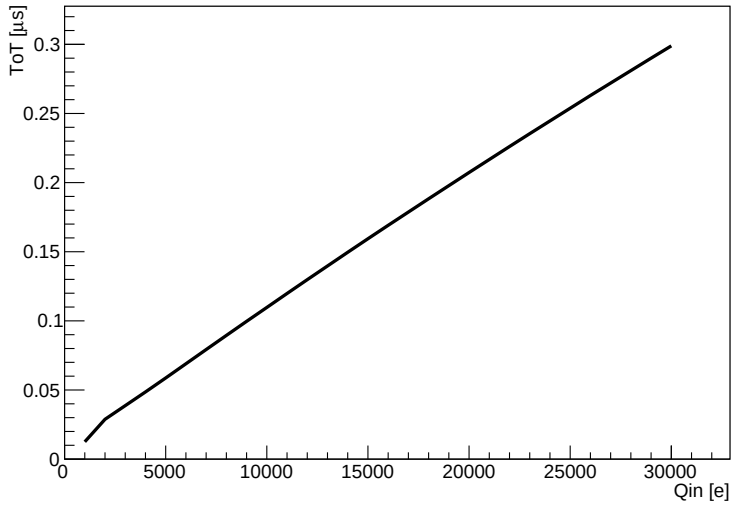

Figure 4.19: ToT as a function of the input charge

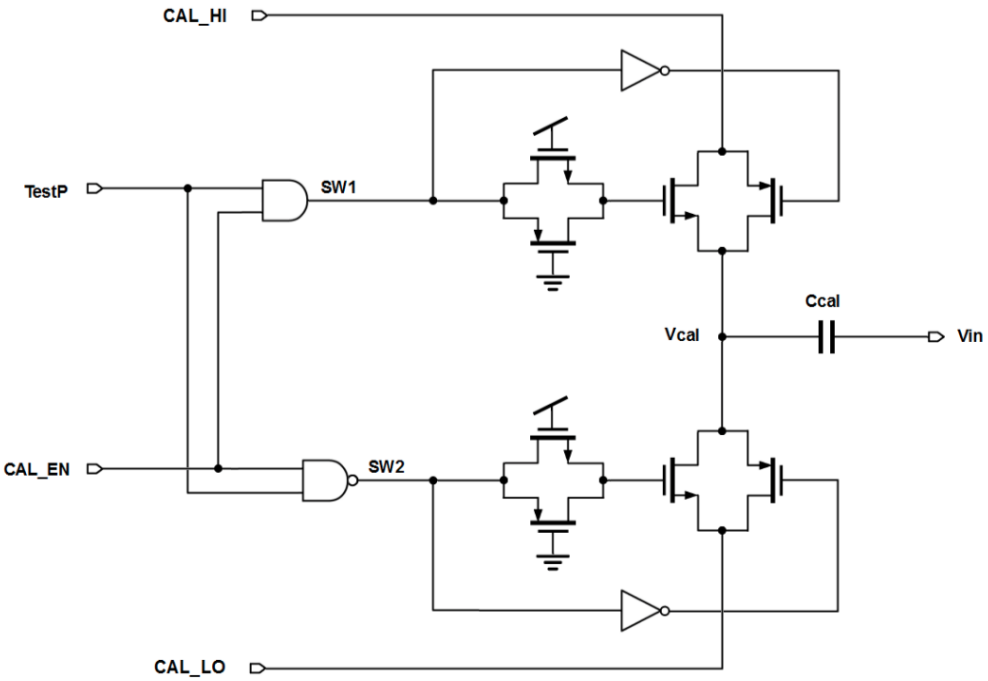

Figure 4.20: Schematic of the calibration circuit

4.3 Discriminator

As shown in figure 4.2, the output of the preamplifier is AC coupled to the discriminator. The output baseline of the CSA is in fact subject to important variations between pixels due to the mismatch effects, as shown in figure 4.21. From a pixel to another it is possible to have a difference in baseline around 30 mV, which corresponds to almost 1 ke$^-$ at the input. These fluctuations can therefore have a significant impact on the effective threshold voltage applied to the discriminator. A capacitive coupling between the preamplifier and the discriminator is enough to get rid of the effect since it is caused by a DC component. The baseline required by the second stage is then restored thanks to the V_{BL} voltage provided externally.

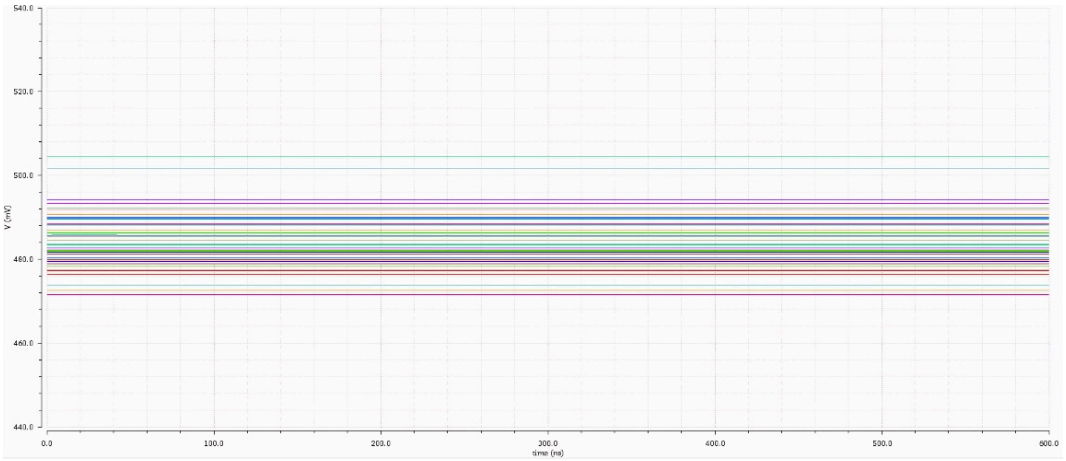

Figure 4.21: CSA output baseline values for 50 different pixels

4.3.1 Differential Amplifier

The first block of the discrimination stage is the Differential Amplifier. Its schematic is illustrated in figure 4.22.
The input of M1 is the signal coming from the preamplifier with the corrected baseline. The same DC level with an additional ΔV_{TH} threshold voltage is instead fed into the gate of M2. This stage produces therefore the differential signal needed for the hit discrimination. In addition, it is designed to provide a small additional gain. Considering that the small-signal conductance seen at the output nodes is given by:

$$g_{ds,out} = g_{ds12} \parallel (g_{m34} + g_{ds34}) \parallel g_{ds67} = g_{ds12} + g_{m34} + g_{ds34} + g_{ds67} \simeq g_{m34} \tag{4.3.1}$$

Device	Size (W/L) [$\mu m/\mu m$]
M1	10/1
M2	10/1
M3	0.3/1.5
M4	0.3/1.5
M5	1.4/2
M6	0.6/1.5
M7	0.6/1.5

Table 4.5: Transistor sizing of the Differential Amplifier

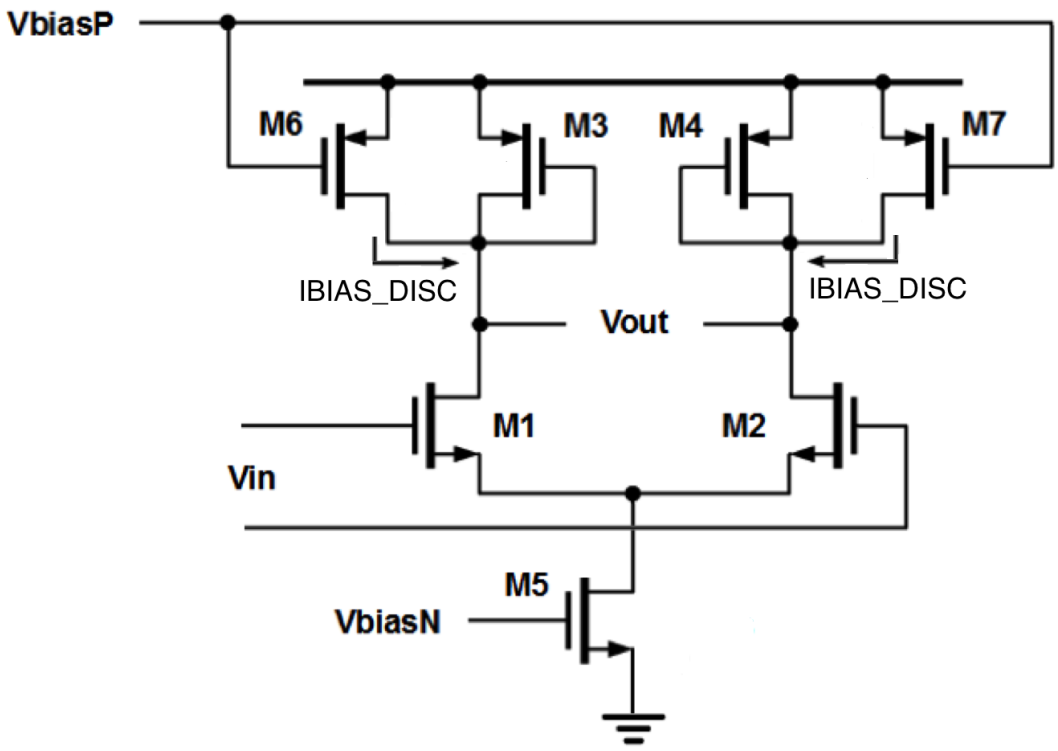

Figure 4.22: Schematic of the Differential Amplifier

The output conductance can then be approximated with g_{m34} since it is around one order of magnitude larger than the other contributions. As a result, the differential gain of this stage can be written as follows:

$$A_D = \frac{g_{m12}}{g_{ds,out}} \simeq \frac{g_{m12}}{g_{m34}} \tag{4.3.2}$$

Therefore, in order to achieve a reasonable gain, the transconductances of M1-M2 should be significantly larger than the M3-M4 ones. For this purpose the current splitting technique, similar to the one used in the CSA, has been implemented. This stage has been designed such that the bias current provided by M5 is 1 μA, which splits equally in the two branches. The VbiasP is chosen so that 400 nA out of the 500 nA flowing into each branch of the DA are taken by M6-M7. Therefore, since the current in M3-M4 is 1/5 of the current in M1-M2, a similar ratio between the transconductances is realized. As a result, an adequate open-loop gain is achieved. Considering only one branch, these choices of transistor sizing and bias current lead in fact to the following values:

- $g_{m1} = 13.6\ \mu S$;
- $g_{ds1} = 17.9\ nS$;
- $g_{m3} = 2.2\ \mu S$;
- $g_{ds3} = 17.3\ nS$;
- $g_{ds6} = 48.2\ nS$.

They confirm then that the contributions of the single output conductance in the value of $g_{ds,out}$ can be neglected. In addition

$$\frac{g_{m1}}{g_{m3}} \simeq 6 \tag{4.3.3}$$

leads to a gain compliant with the expectations.
Regarding the DA, another important aspect to be taken into account is that in real circuits the perfect symmetry between the two branches is not guaranteed due to mismatch. This situation is well described by figure 4.23. In an ideal differential amplifier, in fact, if the differential input voltage $V_{in} = 0$ the perfect symmetry leads also to a differential output voltage $V_{out} = 0$. Nevertheless, mismatch makes this quantity move away from 0 randomly. Therefore it is possible to state that a circuit suffers from a DC offset corresponding to the value of V_{out} when $V_{in} = 0$. It is anyway more practical to make use of the input-referred offset voltage, defined as the input level which forces the output voltage to go to zero. It has to be underlined that the following condition is verified:

$$\mid V_{os,in} \mid = \frac{V_{os,out}}{A_v} \tag{4.3.4}$$

in which A_v is the differential gain of the amplifier. This expression also show that with a high gain the offset may lead the DA to saturate. Therefore it is another reason why A_v should be kept relatively low, usually below 10. In this application, the main limit rising from the offset is the fact that the effective threshold voltage seen by the discriminator will be different between the pixels even if the global threshold voltage is the same. As a consequence, this offset contribution needs to be minimized in order to keep the threshold dispersion under control [27].

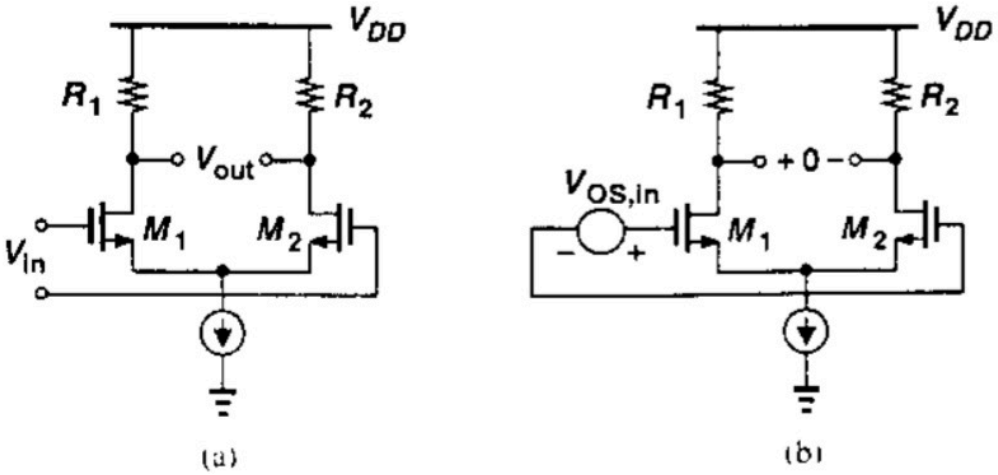

Figure 4.23: Differential pair with offset measured at the output (a) and same circuit with offset referred to the input (b) [27]

Offset compensation

As explained in chapter 3, the straightforward method for offset compensation foresees the implementation of a tuning DAC. Nevertheless, this is not the only possible way of reducing the offset. The choice of a discrete-time comparator, implemented in this design, allows in fact the optimization of the offset through the so-called "autozeroing" techniques. They consists in the usage of sampling switches which are implemented to periodically sense and store the mismatch variations on capacitors such that in normal operation it is subtracted from the input signal. The offset storage capacitor can be placed at the input or at the output of the DA. As figure 4.24 shows, in this application the latter case has been implemented, in the so-called "output offset storage" technique [27].
The DA offset compensation diagram is illustrated in figure 4.23. When the offset compensation phase starts, the Φ_2 switches are opened, so that the DA is disconnected from the first stage.

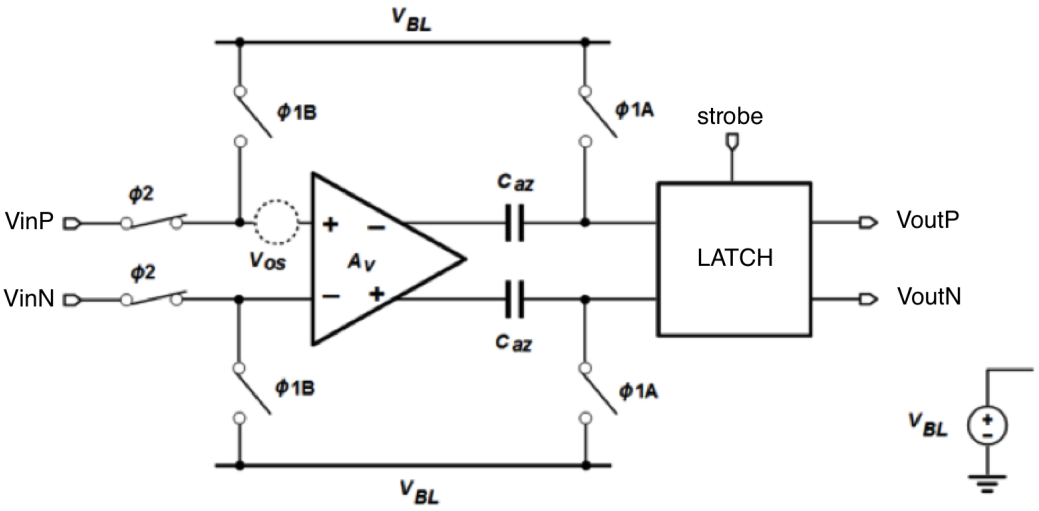

Figure 4.24: Discriminator block diagram

At the same time, the Φ_{1A} and Φ_{1B} switches are closed. In this way, the two inputs of the DA are connected to the same voltage. At the same time, also the two outputs of the C_{az} capacitors are tied to V_{BL}. Therefore, the capacitors see a voltage difference between their two ends corresponding to the value of $V_{os,out}$ given by equation 4.3.4. It is then stored in a given amount of time depending on its amplitude and on the size of C_{az}. Once the storing is complete, the Φ_{1B} are opened with a small advance in order to compensate also residual contributions due to charge injections, which are limited in this technology but still present. Finally, when also Φ_{1A} is opened Φ_2 is closed and the architecture can restart the normal signal processing. This method offers the advantage of a very compact implementation, without the addition of auxiliary blocks. In addition, the pixel-by-pixel trimming procedure needed with the DAC solution in not requested, since each pixel provides by hardware to the offset cancellation once the $\Phi_1 - \Phi_2$ signal pattern is provided.

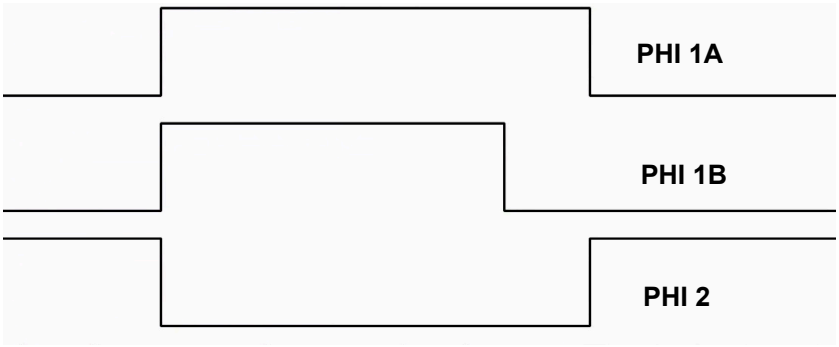

Figure 4.25: Diagram of the offset compensation switches control signals

It has to be underlined that the period in which the offset compensation is performed adds to the dead time of the front-end, since a potential input signal can not be processed. Nevertheless, the LHC particle filling is not totally full of 25 ns spaced bunches. In fact, the 27-km long tunnel contains also a 3 μs abort gap that has to be included in order to properly set up the extraction kicker of the LHC beam dumping system [67]. Therefore every around 86 μs a period of 3 μs without collisions occurs in the CMS detector. As a result, it is possible to take advantage of the abort gap in order to perform the offset cancellation without any increase of the dead time. The capacitors have therefore to be sized such that the offset compensation

frequency can be kept at maximum equal to the one of the abort gap. Once the offset is stored, in fact, the capacitors start to slowly discharge due to the leakage currents of the transistors to which they are connected. In particular, the gate leakage of the two input transistor of the latch stage has a significant influence. These devices are characterized by a constant current around 500 fA. Supposing an offset value stored on the capacitor around 10 mV, the value of the capacitor should be sized so that only a fraction of it is lost, as an example one tenth. As a result:

$$C_z = \frac{I_{gate,leak} t}{V_{os,lost}} = \frac{500\ 10^{-15} A \times 86\ 10^{-6} s}{1\ 10^{-3} V} \simeq 50\ fF \tag{4.3.5}$$

A larger value could have been even better, given that the switches suffer an increase of the source-drain leakage with irradiation, but the area taken by the two C_z capacitors inside the pixel would have been too large. Figure 4.26 shows how the differential voltage taken at the output of the capacitors slowly moves with time. As a result, an offset compensation phase every abort gap period is enough to guarantee a proper operation of the analog readout system.

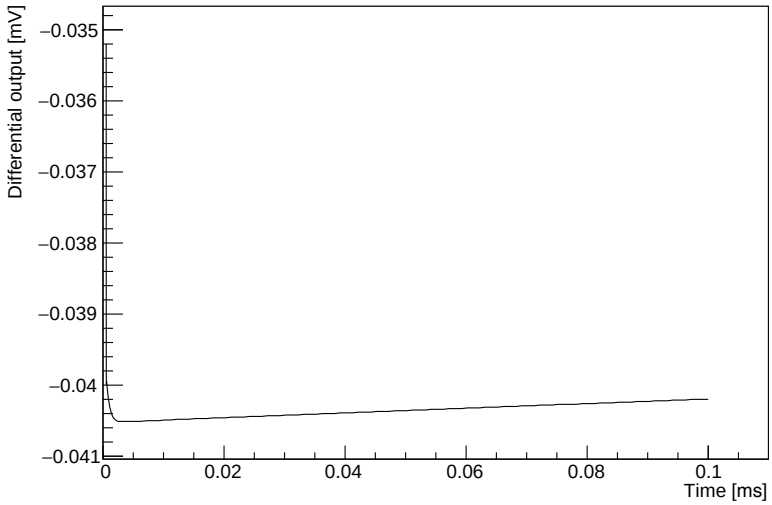

Figure 4.26: Discharge of the autozeroing capacitors with time

An example of the mismatch effect extent between pixels is given by figure 4.27. In this case the differential voltage before the capacitors, and therefore without any compensation, has been simulated for 100 pixels. It corresponds to the effective threshold applied at the latch input in case no offset compensation is applied. Due to the offset, the dispersion is significantly large, with a RMS around 115 electrons. Such a value can not be tolerated. Then the same procedure is applied by measuring the differential voltage at the output of the capacitor. The reduction of the threshold dispersion is huge: it is fact decreases to around 9 electrons RMS, as illustrated in figure 4.28. It confirms therefore that the offset compensation technique adopted in this design works properly and it is a crucial block for the correct operation of the front-end. A last remark has to be made about the switches: they have been designed as CMOS switches, i.e. a NMOS and a PMOS with the same size connected in parallel. This configuration in fact minimizes the amount of the charge injections, since the effects due to the NMOS and the PMOS have the opposite polarity.

4.3.2 Positive feedback latch

The proper discrimination stage consists of the positive feedback latch presented in figure 4.29. As hinted in the previous paragraph, it is a synchronous block. In fact, every time the strobe

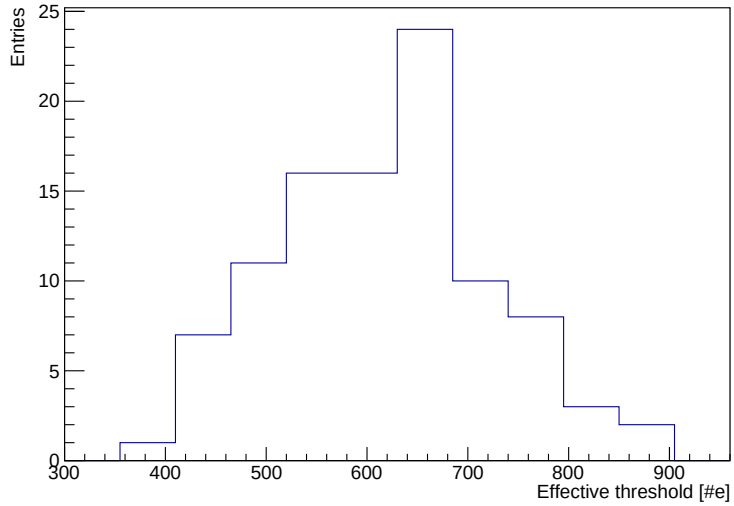

Figure 4.27: Threshold dispersion between 100 pixels with offset compensation

Device	Size (W/L) [$\mu m/\mu m$]
M1,M2	2.5/0.3
M3,M4	0.2/0.06
M5,M6	0.2/0.06
M7,M8	0.6/0.12
M9,M10	0.5/0.06
M11,M12	0.5/0.06

Table 4.6: Transistor sizing of the positive feedback latch

signal, which corresponds to the 40 MHz clock, moves high the latch performs a comparison between the two inputs coming from the DA generating two digital signals, VoutP and VoutN. One of the two will be high and the other low depending on the polarity of the input voltage difference.

This stage is a regenerative comparator based on the principle of positive feedback [28]. It allows to quickly regenerate a small voltage difference to full digital levels. The concept of positive feedback can be simply understood by studying the configuration proposed in figure 4.30. Here two inverters are placed so that the output of one is connected to the input of the other. As a starting condition, $V_x = V_y = V_{DD}/2 = 600\ mV$. In this case the system is balanced because the same current flows in every block. However, this configuration can be easily changed by even small perturbations due to the large small-signal gain of the inverters. If, as an example, V_x increases a bit so that $V_x > V_y$, a larger current flows in the inverter on top, resulting in a further decrease of the V_y voltage. As a consequence, the current flowing in the bottom inverter diminishes, moving the V_x voltage even higher, increasing then the voltage difference between V_x and V_y. This process then continues until V_x saturates to V_{DD} and V_y to ground. At this point, the system reached an equilibrium and, neglecting gate leakage current, no more power is required [65]. As a result, such a system is characterized by power consumption only during transitions.

Among the possible implementations of latched comparators, the one presented in figure 4.29 has been chosen. The input common source transistors sink the current from the regenerative cross-coupled inverters. The operation of this stage is controlled by the *strobe* digital signal.

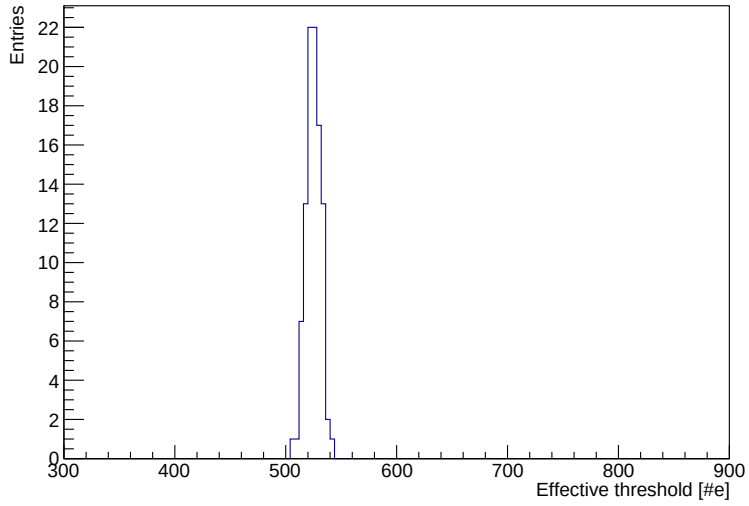

Figure 4.28: Threshold dispersion between 100 pixels without offset compensation

When the latter is low, the latch operates in reset mode and the M9-M10 PMOS transistors are turned on. As a result, both V_x and V_y are tied to V_{DD}. At the same time also the M3-M4 NMOS switches, controlled by the XNOR gate, are on. When the strobe signal moves high the regeneration phase is activated. Therefore, M7-M8 are on while M9-M10 are turned off. At the beginning, the V_x/V_y voltages move down with respect to V_{DD}. Then, depending on the value of the gate voltages of the input pair M1-M2, the positive feedback will move one of the two outputs towards ground and the other to V_{DD}. In addition, as soon as the voltage difference between V_x and V_y starts to be enlarged by the cross-coupled inverters in the regeneration phase, the inverter buffers of V_x and V_y amplify the voltages. As a consequence, the XNOR gate is triggered and switches off the M3-M4 devices. In this way the input pair is decoupled from the output nodes before the voltage variation of V_x/V_y is widened until it reaches the supply voltages. This additional feature allows to limit the differential kickback noise caused by large variations of output nodes since the inputs are decoupled from the outputs [68].
Kickback noise can be in fact an issue in analog front-ends. If no precaution is applied, the large variations of the output voltages are coupled by the gate-drain capacitance of the input transistors back to the input nodes of the latch. In other words, significant spikes are injected back into the Differential Amplifier. If these effects are too large, they can lead to errors.
The implementation of the XNOR gate provides, in addition, another significant advantage. In fact, when the strobe is high and the transition is completed, if the drains of M1 and M5 (and the same applies for M2 and M6) are shorted together, a current consumption path will be present even after the transition is complete. Supposing a case in which $V_x = 0$ and $V_y = 1.2$, this configuration makes M12 on. Since the strobe is high also M8 is on, exactly as M2. M6, in turn, is off because its gate voltage is equal to 0. Therefore, a path for a DC current is open in the M12-M8-M2 and it can lead to an important current consumption which remains until the strobe signal moves low. The XNOR solution solves the problem because when the transitions starts M4 is turned off and the DC current path is cut. In case $V_x = 1.2$ and $V_y = 0$ the same conclusions apply for the left branch. Therefore, the addition of the XNOR gate guarantees also that the comparator takes power only during transitions. As a result, the average power consumption of the block is about 1 μW.

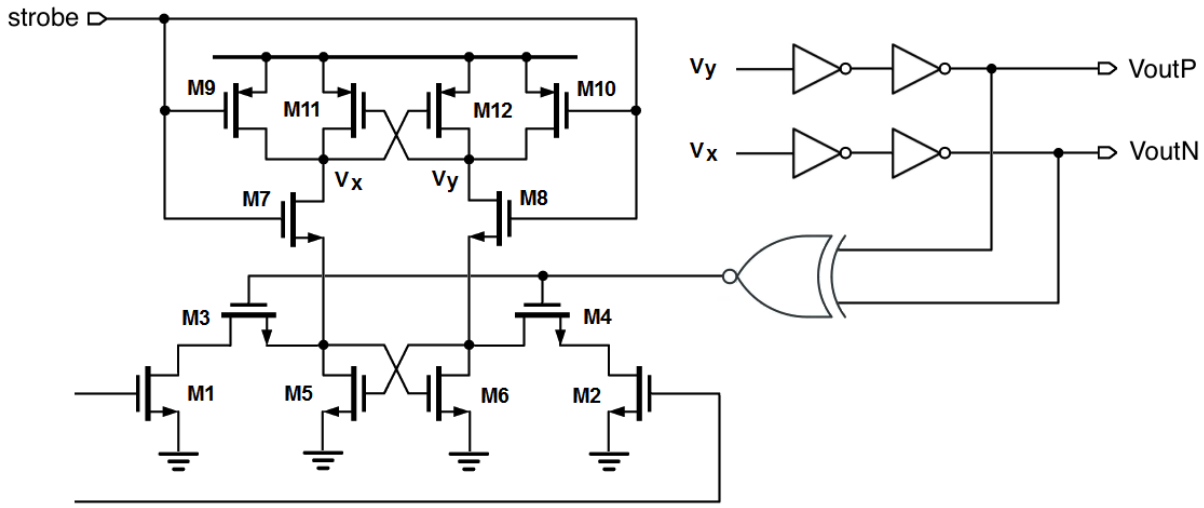

Figure 4.29: Schematic of the positive feedback latch

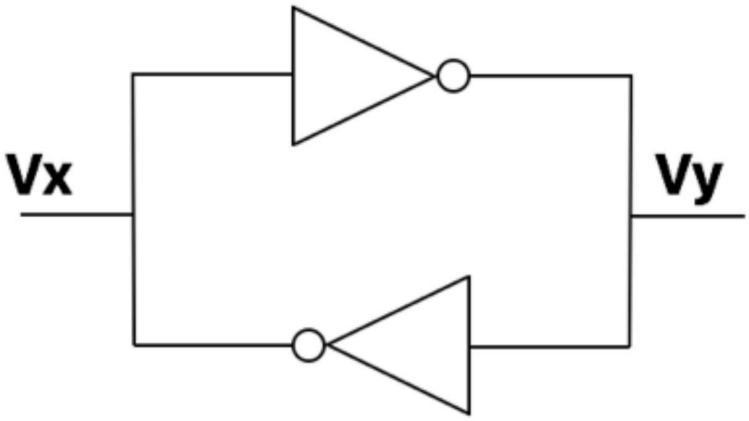

Figure 4.30: Example of positive feedback with a couple of inverters

Latch dynamic offset

The positive feedback latch is subject, as the other blocks, to mismatch. In principle, in fact, this stage is able to detect very small input differences. Few mV more on one of the two input will make V_x or V_y move towards V_{DD}. Nevertheless, also in this case mismatch affects the perfect symmetry of the block. In principle, the equilibrium point in which none of the output diverges to the power rails is given by an input difference equal to 0. Nevertheless, with mismatch it moves of some mV in negative or positive directions from pixel to pixel. As a result, this effect behaves like an additional residual offset which can not be compensated.
In order to simulate this effect, a common mode input voltage equal to 500 mV, corresponding to the baseline of the two outputs of the DA, has been set at the two inputs of the positive feedback latch. Then a ΔV has been applied to the gate of M1 sweeping from -50 to 50 mV with 5 mV steps. Each step has been performed with 200 events in order to have enough statistics. As figure 4.31 shows, with -50 mV all the pixels behave in the expected way: since the gate voltage of M1 is lower than the M2 one, V_x increases and in the positive feedback makes V_x move towards V_{DD} and V_y towards ground. In the figure the latter terminal is plotted. With the progressive increase of ΔV mismatch make some pixels behave like if the equilibrium points has been already passed even if it is still below 0. In other cases, instead, it is located above

0. The RMS of the curve is around 10 mV. Measurements shown in chapter 5 suggested in fact that the original sizing of the devices composing the latch was too small resulting in a too large dynamic offset. As a consequence, at the expense of speed and power, the transistor sizes have been increased in order to improve this crucial figure. A detailed discussion on this part is provided in chapter 5.

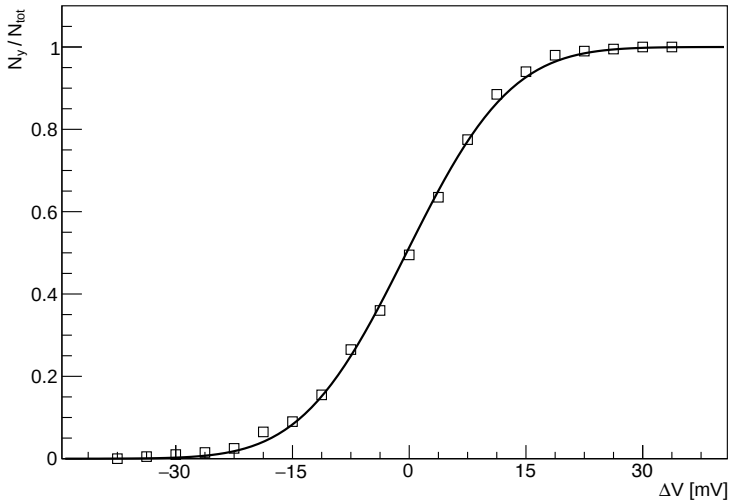

Figure 4.31: Representation of the latch dynamic offset

Fast oscillator for ToT

A key advantage of this implementation of the latched comparator is the possibility of turning it into a fast local oscillator for ToT counting. A similar technique based on the implementation of local Voltage Controlled Oscillators (VCO) has been widely used in HEP applications. The generation of the high frequency clock at the pixel level has in fact allowed to limit the power consumption that would have been significantly increased by the distribution of such a signal across the whole pixel matrix [69]. Nevertheless, the discrete time comparator presented here can act itself as an oscillator, without requiring the implementation of an additional block, by means of an asynchronous logic. This method is frequently used in SAR ADC applications [70].

The logic architecture implemented in this design is presented in figure 4.32. The activation of the fast mode operation is provided by a multiplexer which makes the *strobe* signal switching between slow and fast operating mode. The principle is the following: in normal operation, the latch receives the 40 MHz clock and samples the analog signal around the peaking time. If it is below the threshold, the 40 MHz clock remains active. If instead the signal goes above the threshold the multiplexer switches, turning on the the asynchronous logic feedback loop which makes the latch operating as an oscillator. The frequency of the oscillation can be tuned by means of a voltage-controlled delay line element, indicated by the red circle in figure 4.32. It is based on the principle of current-starved inverters: a current source is added between the PMOS and the NMOS in order to limit the sinking capability of the inverter. The I_{ctrl} branch of the cell is put outside the pixel since it is enough to choose the same current for all the pixels. This technique introduces an asymmetry between pull-up and pull-down transitions. As a result, two starved inverters in cascade are required. The final standard inverter is inserted in order to provide the necessary drive strength for the cell. The choice of providng the control voltage in the current domain comes from the fact that the distribution of a DC voltage is not

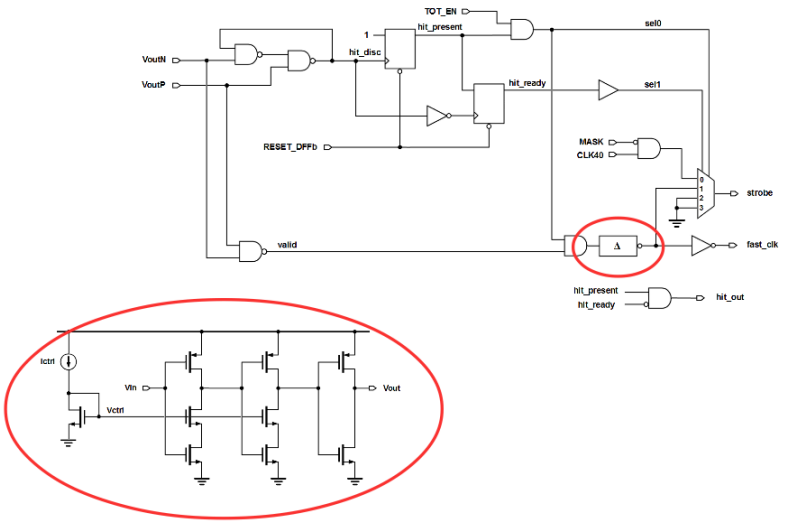

Figure 4.32: Scheme of the hit logic required for the fast oscillation operation

reliable enough. In fact, small variations of the V_{GS} for the starved devices lead to significant fluctuations of their R_{on}, which reflects on the oscillation frequency. Figure 4.33 shows how the frequency changes as a function of the current flowing in the delay line. Although the current required to reach, as an example, a frequency equal to 500 MHz is quite large (around 4 μA) the impact on the average power consumption is lower: in fact it contributes only during transitions, which happen only when a hit is detected. Therefore, the percentage of time in which this current value is reached inside the pixel is small. Finally, once the analog signal goes below the threshold, the latch is moved back into 40 MHz operation [65].
The detailed operation of the logic is the following. The two latch outputs, VoutP and VoutN, are fed into a NAND gate which generates a *valid* signal. When the strobe is low, both outputs are high and the *valid* remains low. In turn, when the strobe moves high and the comparison is performed, one signal between VoutP and VoutN is high and the other is low, pulling the *valid* signal to high. The latter is put into a AND with the output of the topmost branch of the logic. Here the VoutP and VoutN signals are used to feed a NAND-based set-reset latch. If a particle hit is detected, this structure generates a high value for the *hit_disc* signal. Then the latter is put in AND with a TOT_EN digital pulse. This configuration bit allows to disable the fast oscillation logic in case a normal 40 MHz operation is desired also when a particle is detected. Therefore, if TOT_EN and *hit_present* are high the same applies for the *valid*, which is then injected and delayed by the current-starved element. The most important block of the logic is the multiplexer. The 0 input receives the standard 40 MHz clock ($CLK40$). The latter is also put in a AND with the inverse of a $MASK$ signal, which can be used to completely disable the clock in the pixel if for example the latter shows itself as a noisy pixel. The 1 input is instead used to close the feedback loop: it turns the latch into a local oscillator, receiving the delayed *valid* signal and feeding it back to the latch. The other two inputs are used to implement idle states. 2 is used if a binary-only mode is implemented, while 3 is used to stop operations after a hit has been detected if fast TOT has been enabled.
An example of the latch operation with the 40 MHz clock is provided in figure 4.34. Supposing that the input signal injection happens at $t = 0$, after 25 ns, in correspondence to the clock rising edge, the discriminator sees that the signal is above the threshold and starts firing. In this case the fast ToT mode is disabled and the frequency of the discriminator output is 40 MHz. It stops when the signal moves below the threshold. Figure 4.35 shows instead the situation in which the fast clock, tuned at a 200 MHz frequency, is enabled. In this case the $VoutP$ always starts in correspondence of the 40 MHz clock rising edge, because the oscillation

begins only when the analog signal is above the threshold. Subsequently, the latch is driven by the internally generated clock, and the number of oscillations can be counted in order to perform a high-precision measurement of the ToT. Similarly to the standard configuration, the fast clock stops when the analog signal goes below the threshold. Therefore, the impact of power consumption is very limited because the number of transitions is increased only when a input signal is detected. As a result, it happens in around 1% of the time, making the additional power contribution negligible.

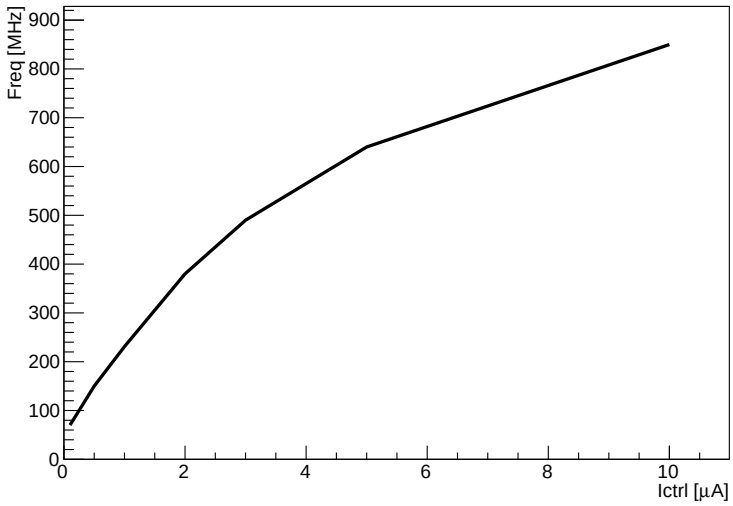

Figure 4.33: Oscillator frequency as a function of the delay line current I_{ctrl}

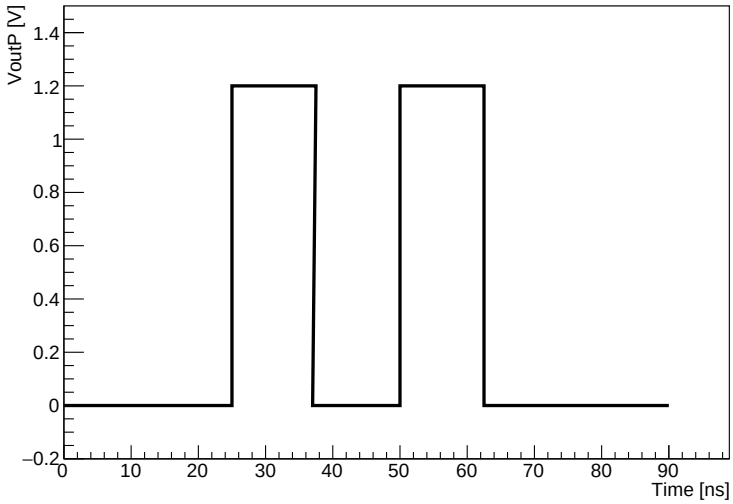

Figure 4.34: VoutP for the latch operating with the 40 MHz clock

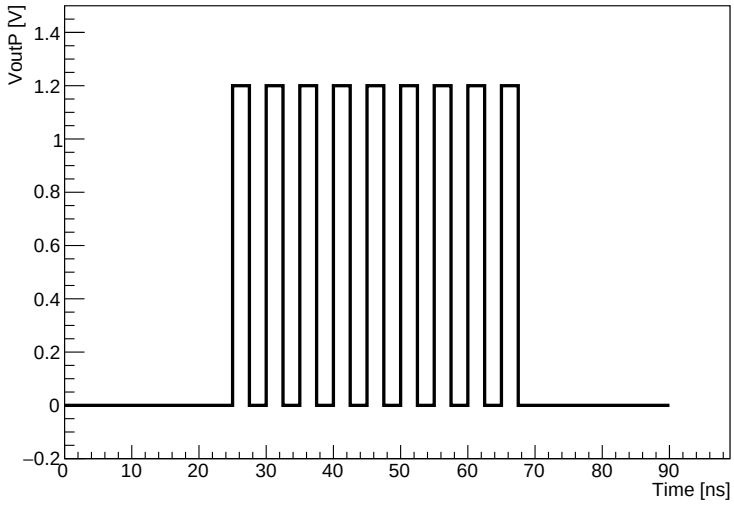

Figure 4.35: VoutP for the latch operating with the fast clock tuned at 200 MHz

4.4 Mixed-signal noise contributions

As shown in the last paragraphs, this front-end contains some blocks which are characterized by a digital-like behavior. In addition, complete readout chip implementations like the CHIPIX65 and RD53A demonstrators presented in chapter 6 feature a significant amount of digital intelligence at the pixel level. When analog and digital circuits operate concurrently on the same chip, in the so-called mixed-signal designs, substrate noise becomes an aspect to be carefully taken into account. These effects can be simply explained by studying the behavior of an inverter. In fact, the change of state at the gate leads to a charge from ground to V_{DD} or vice versa of the internal and load capacitances. As a result, fast current spikes are produced during the transitions [29]. Although the spike is very short, in the order of 50-100 ps, it can lead to significant parasitic current injections, driven by three main mechanisms [71]:

- **Impact ionization**: The electrons accelerated by the electric field can create electron-hole pairs by impact ionization. The latter can then flow into the substrate. Normally, it represents a quite negligible contribution with respect to the others;
- **Source/Drain capacitive coupling**: due to the pn junctions between the source-drain regions and the substrate a current can be injected in the substrate during the switching of a source-drain node;
- **Supply coupling**: A non-zero inductance is present between the bonding wires and the traces on the Printed Circuit Board (PCB) used for chip testing. As a consequence, fast current transients lead to bounces on the supply rails. In particular the ones on V_{DD} produce currents into the substrate through the capacitive coupling between the PMOS nwell and the substrate. On the other hand, the variations on the ground terminal are coupled to the substrate because of the resistance of substrate contacts and of the ground distribution grid included in the chip.

Supply coupling is usually the dominant contribution, in particular in case the substrate is directly biased to the digital ground. These parasitic currents can have an impact on the analog devices, such as the input transistor of the CSA. Among other effects, they can result in slight variations of the bulk potential, which reflect on the threshold voltage, and modulations of the analog ground line, usually used also for the biasing of the substrate in the analog part

of the chip. As a consequence, some precautions have been taken to minimize these effects in this design. Some of them concern the layout, which is presented in detail at the beginning of next chapter:

- Independent lines for analog and digital power/ground have been adopted, in order to separate as much as possible the two domains. In addition, the positive feedback latch and the delay line have been connected, together with the hit logic, to the digital rails in order to minimize the effects of their spikes on CSA and differential amplifier.
- In the layout the blocks in the two different domains have been spaced as much as possible and shielded by guardrings of substrate contacts;
- The implementation of DNW devices in the critical areas has been performed since this technique provides isolation from the substrate. It is proven to be particularly effective for low and medium frequencies;
- In large prototypes, a number of pads in the I/O ring of the chip have been dedicated to couples of analog and digital power/ground terminals, in order to reduce the series inductance contributions of the wire bondings;
- The power and ground grids inside the chip have been designed using the thickest and largest metals available with the aim of reducing their resistance;
- Where possible, decoupling capacitors have been included inside the pixel matrix.

An important topic in mixed-mode systems is represented by noise induced by the clock. In principle, this aspect can have a significant impact on the front-end described in this work, given also the high frequency operation for fast ToT digitization. Nevertheless, in this case the clock is so fast that its leading and trailing edge occur before the preamplifier is able to develop a full signal and will cancel each other. Therefore, the impact of the fast clock on the front-end performance is substantially reduced. At the same time, this effect is further minimized by the choice of a synchronous discriminator. In fact, if the amplifier output is sampled simultaneously with the clock, the fluctuations on the baseline are captured in the same position and manifest themselves as an offset that may depend on the clock frequency, but it is known and can be calibrated-out for a given clock frequency.

Chapter 5

Test prototypes and measurement results

In this chapter the submissions and the measurement results of two small test prototypes are discussed. At the beginning of the chapter, an overview of the layout of the chip sent to the foundry is given. The successive portion of the chapter is dedicated to the measurement results regarding the main front-end parameters like gain, noise, threshold dispersion and fast oscillation for ToT counting. In addition, the results of a X-ray irradiation campaign performed up to 600 Mrad are reported.

5.1 First prototype

A first small prototype chip, called CHIPIX_VFE1/TO has been submitted to the foundry for production in October 2014. It is presented in figure 5.1 and is characterized by a size of $1 \times 1\ mm^2$. It is composed of a matrix containing 8×8 pixels with the addition of some peripheral blocks providing proper current and bias voltages and structures for testing purposes. One of the key aspects in the preparation of the prototype design is the layout of the single pixel, which is presented in figure 5.2. It features the actual design of the devices contained in the analog front-end which is then provided to the foundry. A proper layout of the building blocks is crucial since its implementation introduces parasitic resistances and capacitances which are not taken into account at the schematic level. These additional elements can have a strong impact on the performance of the front-end and therefore have to be minimized, in particular in the most sensitive nodes. Modern simulators allow to extract the layout parasitic contributions, giving the opportunity of a verification of the additional effects before the submission of the design. These tools have been therefore used to perform a progressive optimization of the layout through a study of the parasitic elements in the most critical parts of the front-end. Each pixel features the analog architecture presented in Chapter 4, together with some additional blocks implemented for testing purposes. The complete list of the elements included in the pixel is the following:

- **1. Input capacitors**: since this kind of prototype is too small to be bonded to a sensor, some capacitors have been added at the pixel input in order to be able to emulate its effects. In particular three capacitors of value 21.5, 43 and 86 fF have been included. The smaller one is hard-wired while the other two capacitors can be disconnected from the input through a switch controlled by an external signal. As a result, studies of the main parameters like noise as a function of the input capacitance have been performed. In the pixel layout also the structures required for the bump bonding have been included, in order to simulate its parasitic contributions even if no sensor is connected. Simulations have shown that it results in an increase of input capacitance in the order of 25 fF;

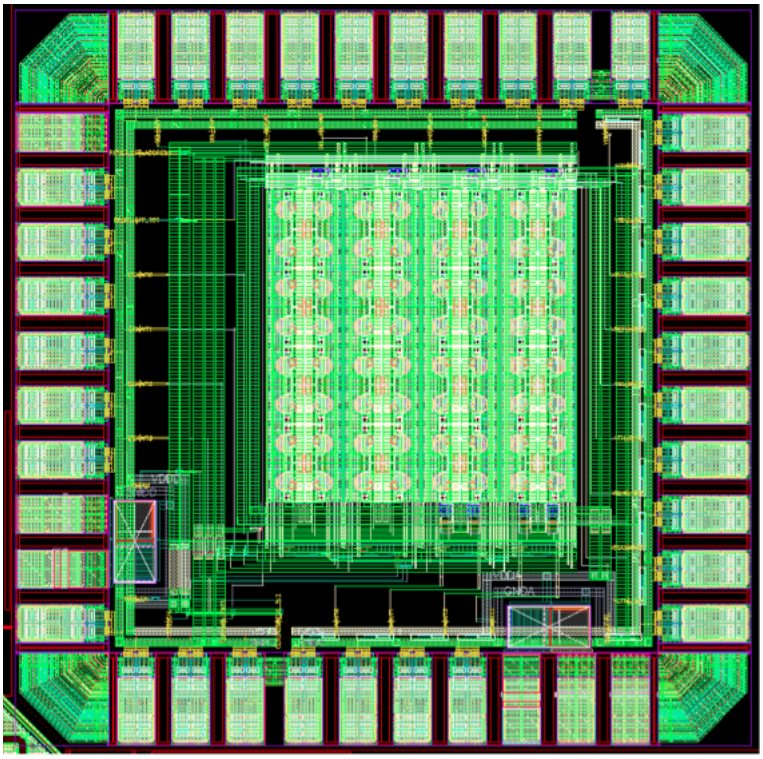

Figure 5.1: Layout of the first chip prototype

- **2. Analog buffer**: in a test chip the goal is to monitor as many quantities as possible in order to have the best characterization of the front-end with respect to CAD simulations. The output of the CSA is of particular interest, since it gives information about the signal amplitude, the constant current discharge for ToT and the real signal shape. Nevertheless, a direct connection of the CSA output node with a pad and, therefore, an oscilloscope probe is not feasible, due to the high capacitance introduced by the latter device. In fact, it would lead to an important instability at the CSA output node, totally comprosiming the signal processing. For this reason, an analog buffer implemented as a simple source follower stage has been included in each pixel. It provides the required impedance matching while preserving the performance of the main analog chain. A switch has been also included so that the buffer can be disconnected when the analog output is not monitored. In fact, even with this technique a few percent degradation of the amplitude and peaking time is expected. A second analog buffer, always based on the source follower principle, has been added in the chip periphery. The choice of a double buffer with progressively larger dimensions is in fact the best compromise in terms of impedance matching, considering that the CSA output capacitance is in the order of tens of fF while a typical probe capacitance is few pF. The chip periphery features also a digital buffer, needed for a proper readout of the discriminator outputs. In fact, in case of a direct connection with the probe capacitive coupling the digital signal is modified, in particular at the high frequencies produced in the oscillation mode;

- **3. Calibration circuit**: the 3a part contains all the digital logic required for the correct operation of the block. The use of this kind of cells allows to obtain a very compact solution of this block. The 3b part basically contains the injection capacitance, implemented with a Metal-Oxide-Metal (MOM) solution in order to have a high precision;

- **4. CSA**: it represents one of the most important blocks from the point of view of the layout. In fact, parasitic elements have an impact on the amplitude and rise time of the

stage. Given its large width, equal to 8 μm, the input transistor has been fingered in order to fit it in the pixel. In addition, all the NMOS devices have been included in a DNW, in order to provide better isolation form the substrate to limit additional noise contributions. Nevertheless, as figure 5.3 shows, a certain deterioration of the CSA performance has to be expected even if the adequate precautions are taken. In this example, obtained in a simulation performed with a 5 ke^- input signal, the amplitude decreases by around 13% and the peaking time roughly 20%;

- **5. Krummenacher feedback**: it is located close to the CSA in order to simplify the routing between the two blocks. The most of the area is taken by the dumping capacitor. Since its value has to be around 500 fF, it has been implemented with a PMOS transistor. In fact, it provides the largest possible capacitance with the minimum area consumption with respect to metal-to-metal solutions. On the other hand, it is not precisely controlled since it depends on the gate voltage which is subject to variations, leading to few percent fluctuations of the capacitance value. Nevertheless, in this node it is not crucial to have a fixed value of the capacitance, it is enough to choose it adequately large.

- **6. AC coupling**: the 6a block is the AC coupling capacitor, this time implemented with a MOM capacitor in order to have very limited variations between pixels during operation. 6b is a transistor used like a resistor for the setting of the Differential Amplifier baseline. Measurements and subsequent simulations have nevertheless shown a RC behavior which determined variations of the baseline. As a consequence, this resistor has been replaced by a simple CMOS switch in the updated versions of the front-end;

- **7. Differential Amplifier**: also in this case the NMOS transistors contained in the block have been put inside a DNW for noise minimization. This block is quite affected by parasitic elements, similarly to the preamplifier. Therefore, the differential gain is expected to decrease of around 20% compared to the schematic level case;

- **8. Offset compensation capacitors**: both are designed as MOM devices in order to make them as similar as possible. Discrepancies between the sizes of the two capacitors would in fact result in a non homogeneous charge/discharge phase, with an increase of the residual offset;

- **9. Offset compensation switches**: blocks needed for the generation of the Φ_{1A}, Φ_{1B} and Φ_2 control signals;

- **10. Positive feedback latch and delay line**: this section contains the parts of the front-end that are polarized in the digital domain in order to reduce the noise deriving from the high switching activity on the analog power. Only the hit logic needed for the oscillation for the fast ToT counting is moved in the digital part of the pixel, since it is a fully digital block. In this first submission latch and delay line have been surrounded by guard-ring to provide better isolation. Nevertheless, an improved pixel area management in the updated versions of the layout has allowed to include this part in a dedicated DNW, so that the analog-powered part of the front-end experiences the best isolation from the noise point of view.

Outside the matrix, other blocks have been introduced. Along with the analog and digital buffers already discussed, the most important elements are the bias cells, which are designed to provide the voltages and currents needed for the correct operation of the pixel. These blocks are based on the current mirror principle. An example is given by the Krummenacher feedback bias cell presented in figure 5.4. The bias branch of the current mirror is therefore located in the chip periphery and receives the current from the PCB. The gate is then connected to the input of the second branch of the mirror located inside the pixel. The size of M1 is equal to

the one of the device in the pixel in order to optimize the matching. A similar scheme is then used for the other currents. Every block is used for biasing 16 pixels. As a result, four blocks of each flavors are contained in the chip periphery.

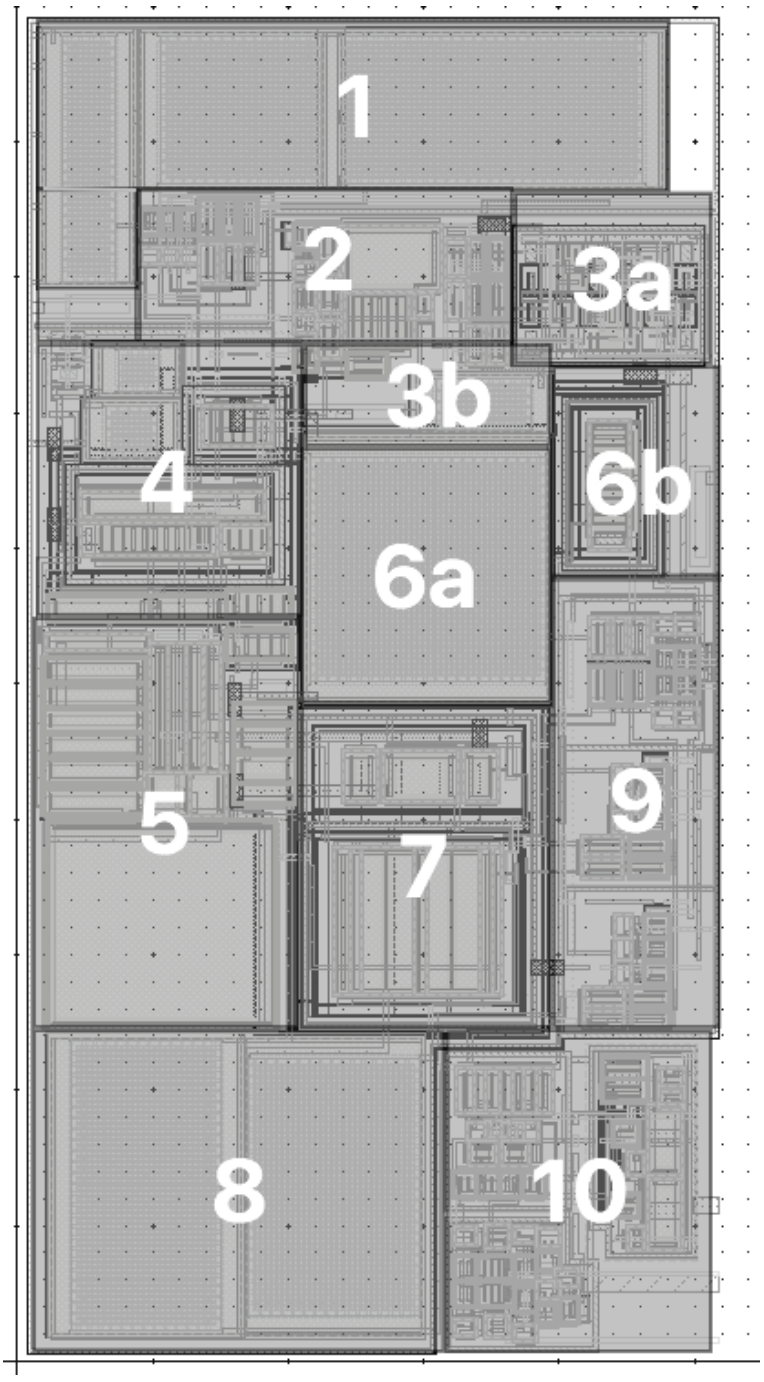

Figure 5.2: Layout of analog front-end included in the CHIPIX_VFE1/TO prototype

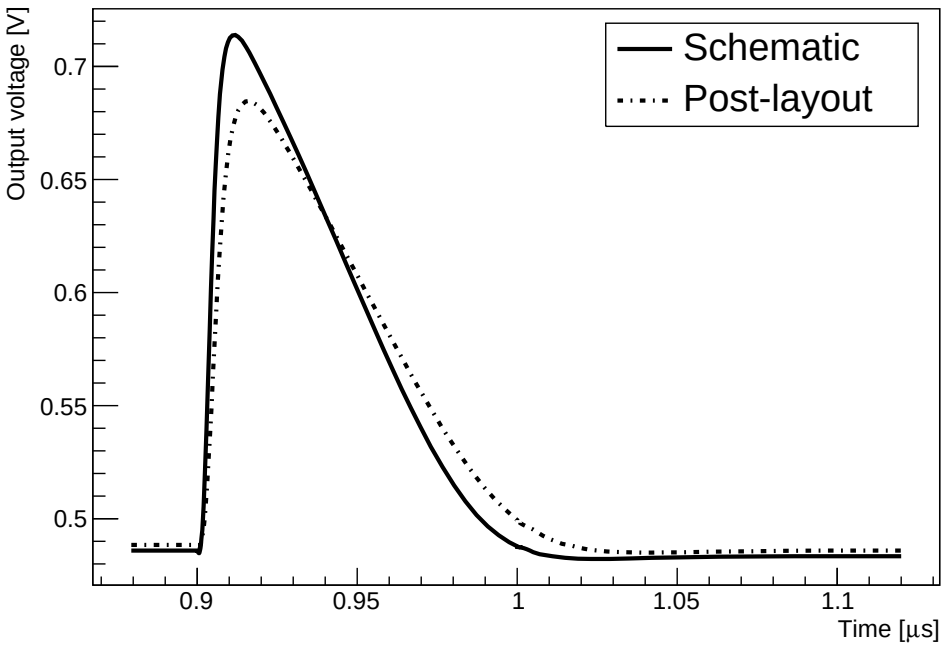

Figure 5.3: Comparison between the response of the CSA to a 5 ke$^-$ input signal at the schematic and post-layout level

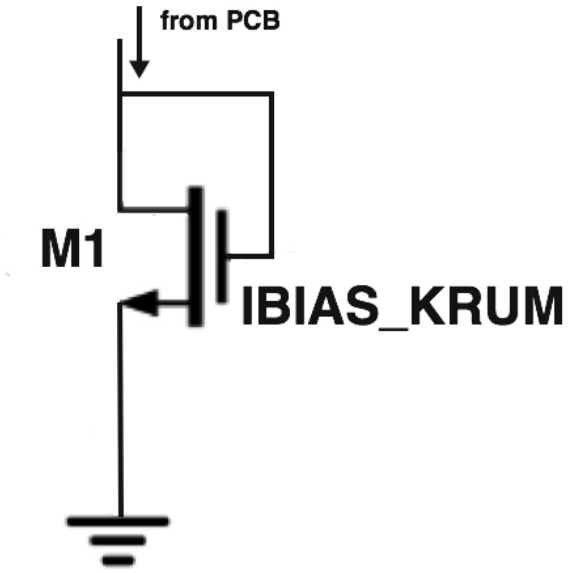

Figure 5.4: Bias block for the Krummenacher feedback circuit

5.2 Test setup

In order to perform the prototype testing, the PCB presneted in figure 5.5 has been designed by INFN Torino. Figure 5.6 shows instead the chip wire-bonded to the board. The latter provides mechanical sustain, wire-bonding landing pads, power, bias circuits, configuration bits and necessary input/output access points [65]. The board is designed to provide separate analog and digital power as required by the chip. Two external power connectors are used to power the board at a nominal supply voltage of 5 V. Since the chip works with a supply voltage equal to 1.2 V, commercial voltage regulators are included to provide it. The board features the external voltage dividers required to complete the biasing scheme. These blocks contain in fact some trimmers which allow to modulate the currents in a specific range ceneterd around the default values listed in chapter 4. The currents flowing into the chip pads are 4 times larger than the nominal pixel values since they have to fill in the 4 bias cells of each type included inside the chip.
From the point of view of the pixel configuration, an 8-contacts slide switch has been implemented to control the global configuration bits needed for the analog front-end:

- **SEL_CIN_50 and SEL_CIN_100** are used to determine the value of the input capacitance. As explained before, their real value is a little smaller: the "50 and 100 fF" have been chosen for simplicity;
- **SEL_C2F and SEL_C4F** allow to adjust the feedback capacitance and therefore the preamplifier gain;
- **ANAOUT_EN and DIGITAL_EN** give the possibility of enabling/disabling the analog and digital buffer outputs respectively;
- **TOT_EN** is used to switch the positive feedback latch clock from 40 MHz to the fast mode;
- **SUBSTRATE_NOISE** allows to enable a ring oscillator located in the digital part of the pixel. Its purpose is to have a further check if a high-frequency free running digital cells increases the noise in the analog part.

Lastly, test points are provided in the PCB in order to have access to the analog and digital output lines printed on the board and connected to the chip pad through the wire bonding.
Another fundamental part of the measurement setup is the Logic State Analyzer (LSA). This instrument has been used to generate the required digital signals which have been fed into the board and then into the chip. In particular they are:

- **CLK40**: the 40 MHz clock needed for the normal sampling frequency of the synchronous discriminator;
- **PHI_CAL**: it is the signal used to start and finish the offset compensation period which is used by the chip to generate Φ_{1A}, Φ_{1B} and Φ_2;
- **RESET_DFF_TOT**: resets the flip-flops implemented in the fast ToT loop;
- **TESTP**: is the signal which controls the injection of the input signal in the pixel. In particular its falling edge provides the signal of the desired polarity.

The whole data acquisition has been performed by using a 4 Gb/s oscilloscope, equipped with an active differential probe which has allowed an optimized sampling of the high-frequency signals.

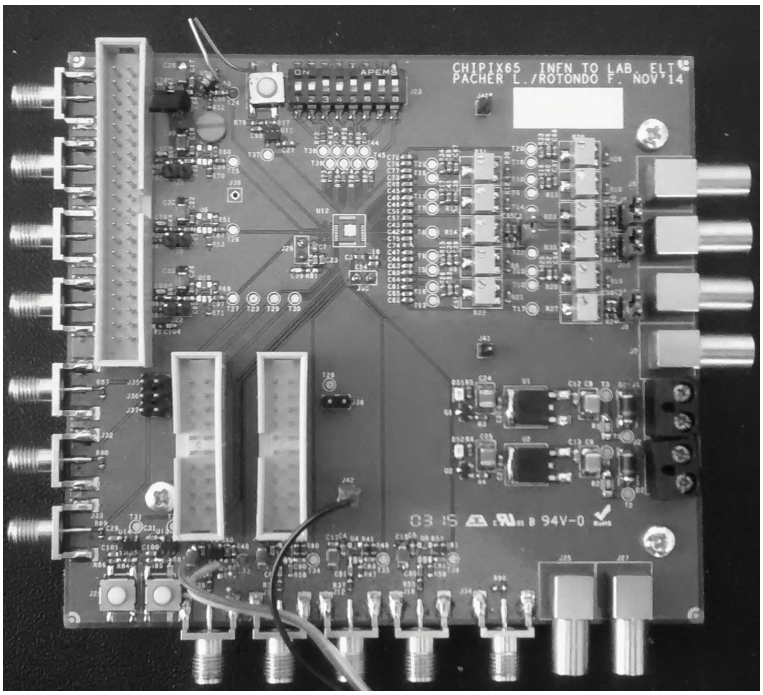

Figure 5.5: Test board

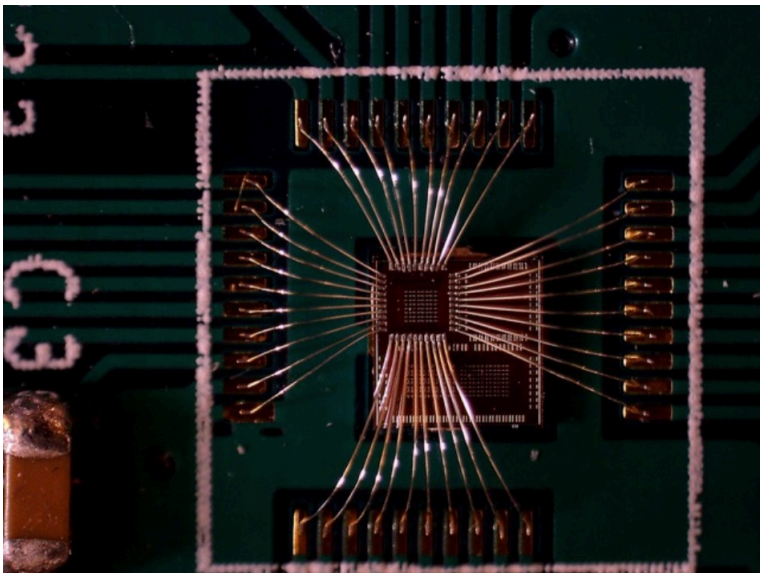

Figure 5.6: Chip wire-bonded to the test board

5.3 Measurements

At the beginning of data taking the default values for the currents have been set via the bias trimmers on the PCB. In addition, the basic pattern of configuration bits has been chosen. The 50 fF input capacitance has been enabled, together with 4 fF for the feedback one. Regarding the offset compensation, a 1 μs-long PHI signal has been provided at the beginning, as an initial condition required for the correct operation of the chip.

5.3.1 Preamp

The preamplifier output signal has been studied in detail. Figure 5.7 shows the shape of the output signal in different configurations of the Krummenacher feedback current for an input signal of 10 ke$^-$. It confirms the expectations derived from the simulations. In fact, in the case with $I_{feed} = 40\ nA$, the time duration of the signal is around 100 ns while with $I_{feed} = 10\ nA$ it is around four times larger. A current of 40 nA allows therefore to maintain the dead time due to the ToT processing well below 1% even in the 3 GHz/cm^2 hit rate case. In addition, the shape of the signal is compatible with the simulations: it features in fact a fast rise time followed by a linear discharge [72].

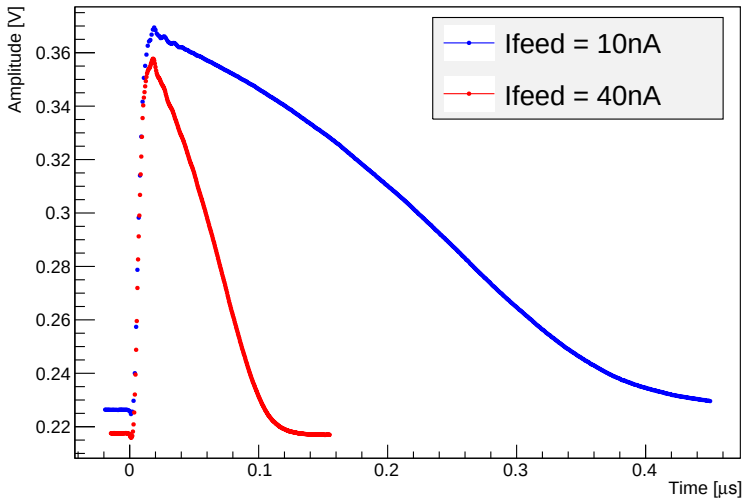

Figure 5.7: CSA output signal for a 10 ke- input charge with different values of Krummenacher current

At this point, the preamplifier gain has been measured for all the 64 pixels contained in the matrix. The input charge chosen for this study is 6 ke$^-$, since in this case the CSA is still in the linear range before starting the gain saturation. In absolute value the gain measured at the oscilloscope is partially attenuated by the two analog buffers, in the order of 30%. Nevertheless, this measurement gives an indication of the gain fluctuations between pixels due to mismatch. As expected, the RMS of the distribution shown in figure 5.8 is only equal to 2.2% of the mean value. Therefore, the impact of mismatch on the preamplifier gain is limited, giving rise to a reasonable uniformity of this quantity across the matrix [72].
The analog signal shape reproduced at the oscilloscope has been also used to determine an "analog" value of the ToT. In fact, a threshold voltage equal to 600 e$^-$ has been added on the screen and the time spent by the signal above the threshold has been calculated. A low feedback current, 10 nA, has been chosen, because longer times made easier the data taking. In fact, the main purpose of this measurement is sto show the linearity of the ToT over the

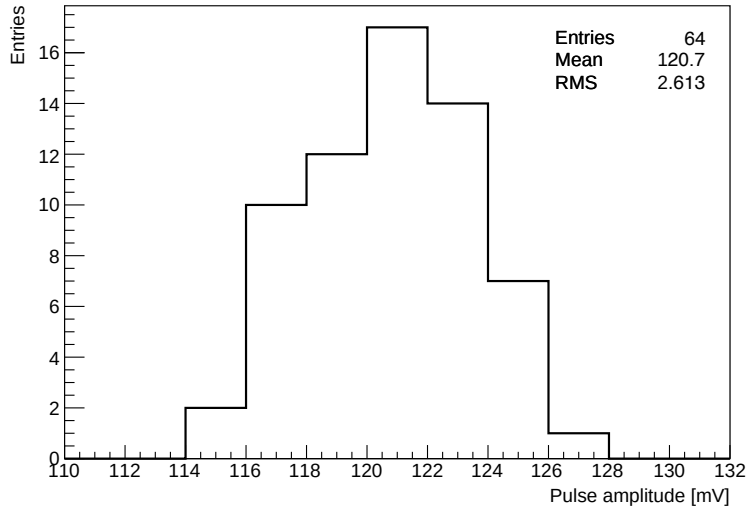

Figure 5.8: Gain distribution across the 64 pixels

whole range of interest of the input charge. Figure 5.9 shows that this feature is verified, with the small deviation form linearity already obtained in the simulations.

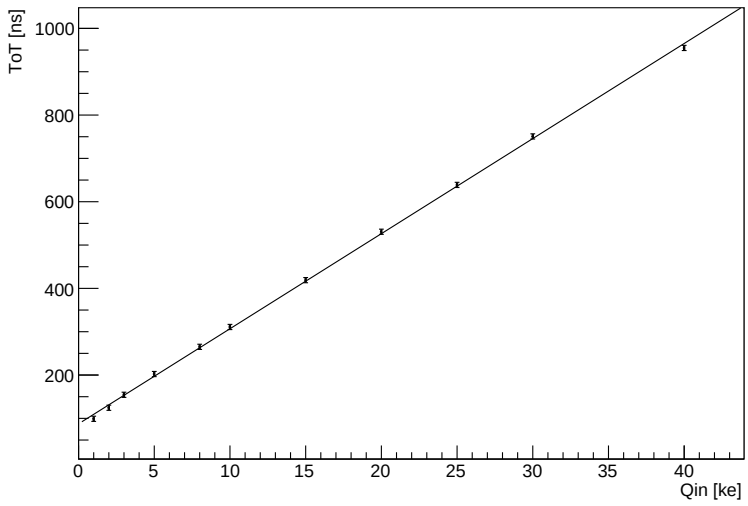

Figure 5.9: Time-over-Threshold linearity measured on the analog signal with a $I_{krum} = 10\ nA$

5.3.2 Discriminator

The following testing phase has been focused on the behavior of the discriminator, taking advantage of the possibility of studying the digital output at the oscilloscope. A proper study of this signal gives in fact the information about the noise and threshold dispersion of the analog front-end. For this purpose, the so-called S-curve technique is used. This characteristic can be obtained in two ways. The first one consists in fixing a value for the threshold and then varying the input charge in a specific range. Considering as an example a threshold value equal to 2 ke$^-$, the trend shown in figure 5.10 can be reproduced. Starting with low charges, like 1.5 ke$^-$, the number of input signals which produces a hit is low, and it progressively increases

with the charge, up to a 100% efficiency around 1.5 ke$^-$. In principle, in an ideal system, the efficiency should remain 0 until the signal is below the threshold becoming then abruptly 1 when the threshold is overcome. Nevertheless, noise makes the transition smoother, leading to this typical S-shaped trend which is described by a Gauss error function. In particular, the required information can be found by fitting the data with the following law:

$$y = \frac{1}{2}\left[1 + Erf\left(\frac{x-\mu}{\sqrt{2}\,\sigma}\right)\right] \tag{5.3.1}$$

μ is the point with 50% efficiency and identifies the effective threshold value, while the sigma is the RMS noise.

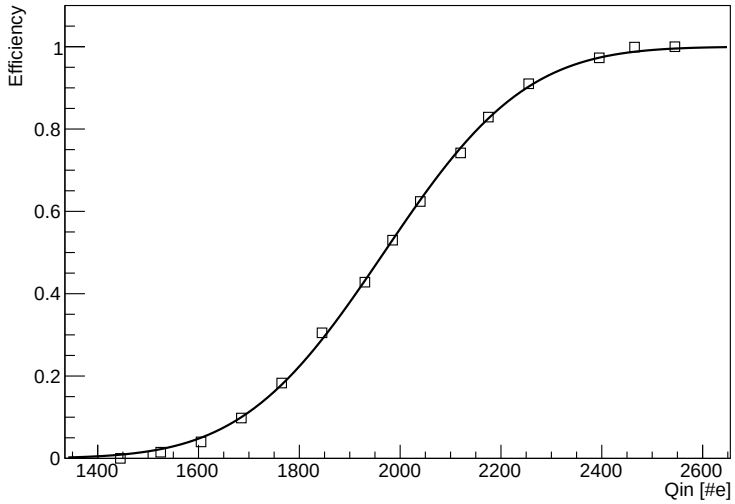

Figure 5.10: Example of a S-curve obtained with an input charge sweep

Alternatively, the same result can be obtained by fixing the value of the input charge while varying the threshold. The resulting S-curve is presented in figure 5.11. The fitting expression is basically the complementary function with respect to the previous case:

$$y = \frac{1}{2}\left[1 - Erf\left(\frac{x-\mu}{\sqrt{2}\,\sigma}\right)\right] \tag{5.3.2}$$

During the most of data taking, the threshold sweep procedure has been applied because the setup allowed a better sensitivity in threshold values. The characterization of a group of 16 pixels has been performed for different values of input capacitance and feedback current, to obtain a proper evaluation of the noise. The plot of the ENC as a function of the input capacitance is illustrated in figure 5.11. Also in this case the two Krummenacher current values used to characterize the analog signal shape have been chosen. The figure shows that the expected linear trend of the ENC is verified in both configurations. At the same time. it confirms that a larger feedback current results in a higher noise figure. Nevertheless, in the region around a sensor capacitance equal to 50 fF the increase in noise is limited to around 20%, reaching a value which is still compatible with the RD53 requirements.
The S-curves have been then used to derive the information about the threshold dispersion. In particular, the $I_{feed} = 40\ nA$ and $C_{input} = 50\ fF$ has been chosen. Firstly, it is important to understand how large the threshold dispersion is in case no offset compensation is performed. Since a correct operation of the front-end requires an initial compensation phase, the choice has been to take the S-curve data 1 ms after a compensation period, when the capacitors

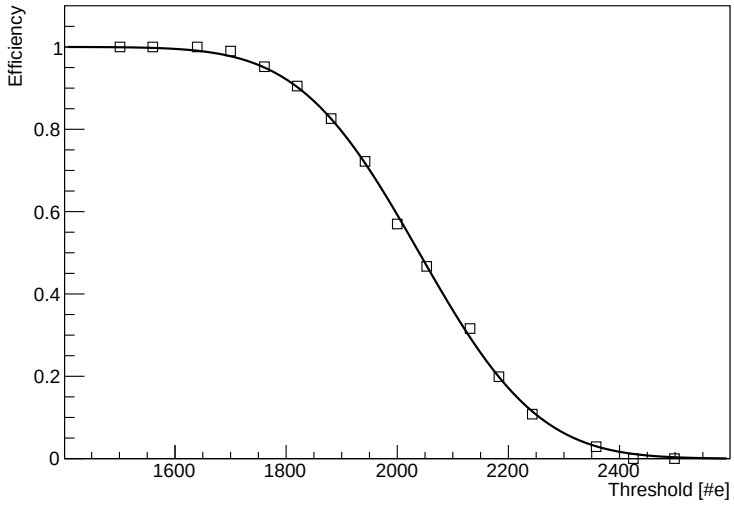

Figure 5.11: Example of a S-curve obtained with a threshold voltage charge sweep

are almost discharged and the resulting configuration reproduces an almost uncompensated front-end. The distribution obtained in this case is illustrated in figure 5.13. It shows that the threshold RMS is equal to 272 e^-, which is an unbearable value compared to the requirements. It therefore justifies the necessity of an offset reduction architecture.

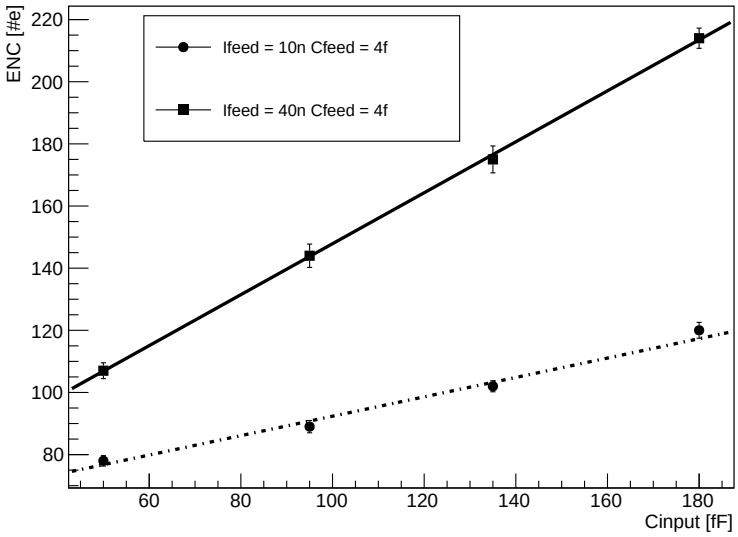

Figure 5.12: ENC as a function of the input capacitance for two values of Krummenacher feedback current

Subsequently, a precise offset compensation pattern has been applied. The frequency has been determined based on the machine abort gap described in chapter 4. Hence, a 100 μs interval has been chosen. In addition, measurements have shown that a 25ns-long compensation is enough. At this point, the S-curves have been measured for 16 pixels and the resulting distribution is shown in figure 5.14. While it shows a substantial improvement with respect to the uncompensated case, proving that the mechanism is working, a RMS equal to 173 e^- is still too large with respect to the requirements [72]. Successive additional simulations on the

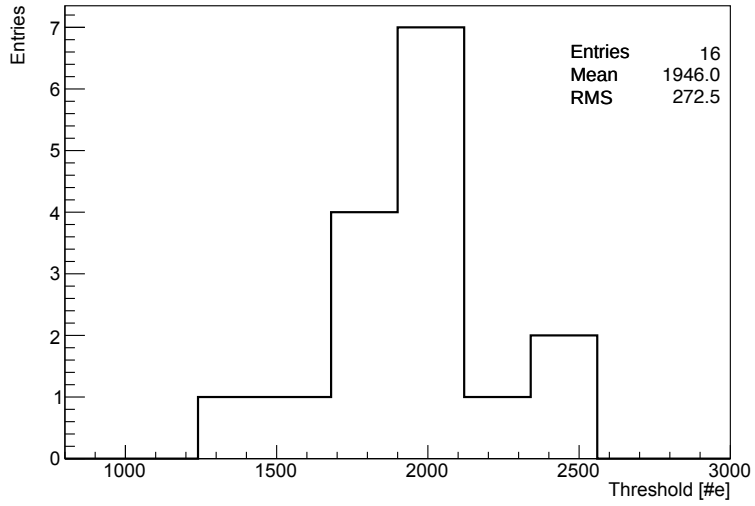

Figure 5.13: Threshold distribution for 16 pixels without offset compensation

front-end have shown that this situation has been caused by two factors:

- A large fluctuation of the DA gain between pixels due to mismatch. Since the threshold voltage is applied at the input of the DA, it amplifies the difference between the signal baseline and the threshold by its gain factor. As a result, important variations of the gain result in large fluctuations of the effective threshold applied at the input of the positive feedback latch. A simulation of this effect involving 100 pixels is presented in figure 5.17. A RMS equal to 6.7 mV, corresponding to around 70 electrons, is obtained. This effect can not be compensated by the capacitors, which take care only of the DC offset contributions and has therefore to be reduced;

- The original sizing of the positive feedback latch, proposed in chapter 4, results in a too large dynamic offset due to mismatch.

These considerations have driven the necessity of producing a second prototype with the aim of a significant improvement on threshold dispersion.
The digital output has been then used to study the fast oscillation mode. An oscilloscope screenshot is reported in figure 5.15 in order to show that it works properly. In fact, once the preamplifier signal moves above the threshold, the oscillator starts and continues to operate until the analog signal is finished. At this point, it stops and the number of oscillations gives the information about the ToT value. In this example the frequency has been kept at around 100 MHz since the setup limited a perfect reproduction of higher frequency signals on the oscilloscope. Nevertheless, it is proven to be working up to at least 500 MHz. In addition, the frequency has been measured for 16 pixels in order to check the fluctuations induced by mismatch. Figure 5.16 shows that the RMS of the distribution is only 1.8% of the mean value. As a consequence, a very good uniformity of the ToT frequency across the matrix is verified.

5.3.3 Summary

The measurement campaign on the first prototype has shown very promising results. In all cases a very good compatibility with CAD simulations have been proven. This is a very important aspect, because it validates the simulation flow giving confirmation of the reliability of their results. Very good results about the preamplifier gain and noise, compliant with the

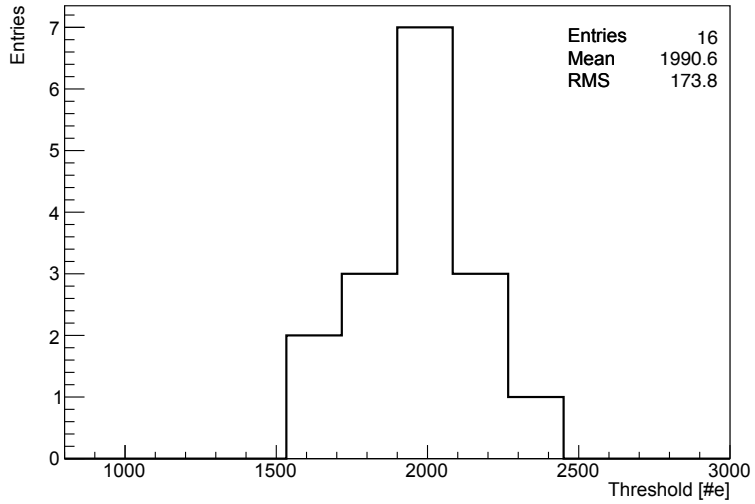

Figure 5.14: Threshold distribution for 16 pixels with offset compensation - First prototype

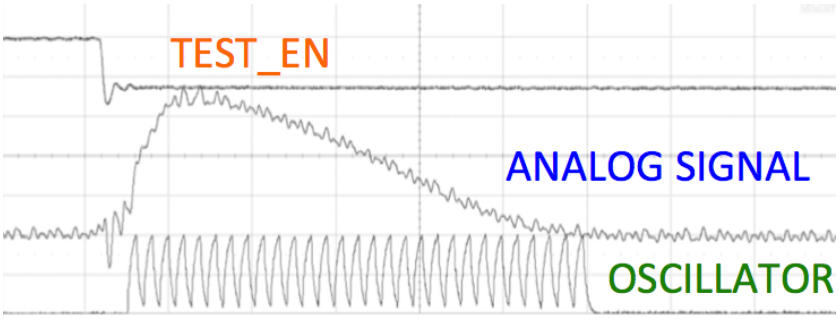

Figure 5.15: Discriminator output in oscillation mode

specifications, have been obtained. In addition, all the functionalities included in the front-end work, in particular the local oscillator for fast ToT counting. Also the offset compensation technique has proven to be working, albeit needing significant optimization. Finally, some measurements have been performed enabling the ring oscillator for substrate noise verification. They have shown no significant variation in the ENC, giving confirmation that high frequency components have little influence on the analog front-end performance.These first encouraging results have driven the submission of a second prototype, described in the next paragraph, with the aim of improving in particular the threshold dispersion figure.

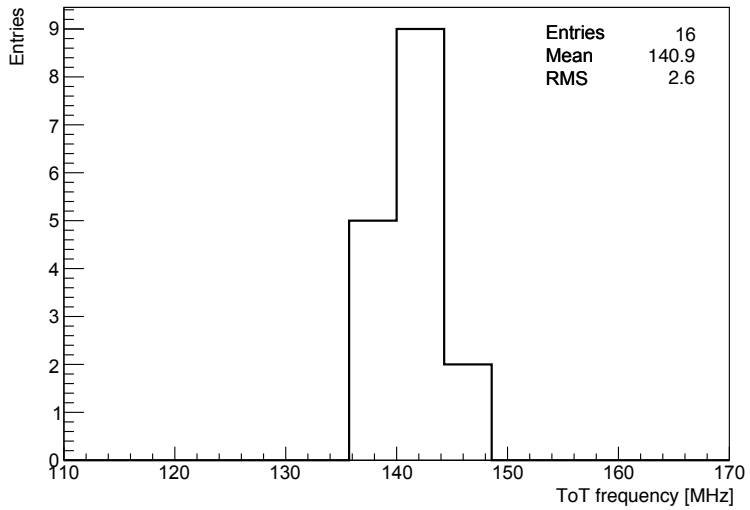

Figure 5.16: ToT frequency distribution across 16 pixels

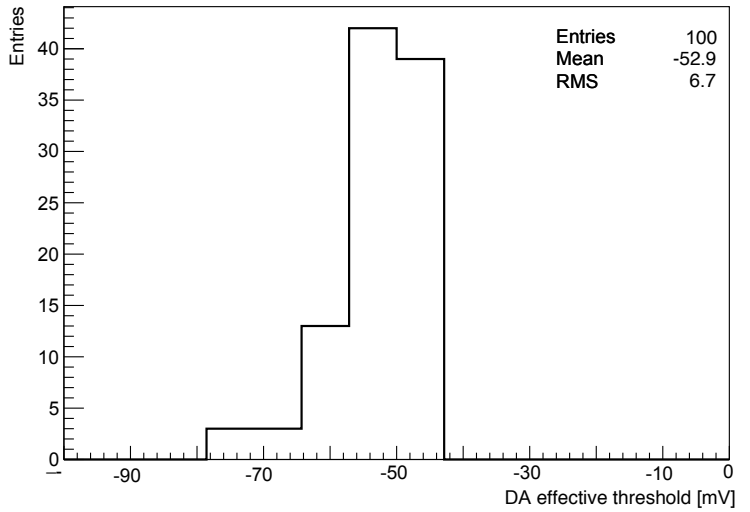

Figure 5.17: Monte Carlo simulations of the DA effective threshold for 100 pixels

5.4 Second prototype

The second prototype, called CHIPIX_VFE2/TO, has been realized with the same size ($1 \times 1\ mm^2$). Since the purpose of this test chip is to check some improvements for the analog front-end, all the unchanged components have been maintained, saving quite a lot of time in the sign off phase. In fact, the pad frame and the chip periphery containing the bias cells and the analog buffers is exactly the same used in the CHIPIX_VFE1/TO chip. Therefore, the global aspect of this prototype is equal to the one shown in figure 5.1. At the same time, significant changes have been performed in the analog section of the pixel, at schematic level first and then in the layout.

5.4.1 Differential Amplifier

Monte Carlo simulations have shown that the largest contribution to gain variations in the Differential Amplifier are ascribable to the current splitting M6-M7 PMOS shown in the schematic in chapter 4. Therefore, the schematic has been at first modified as illustrated in figure 5.18.

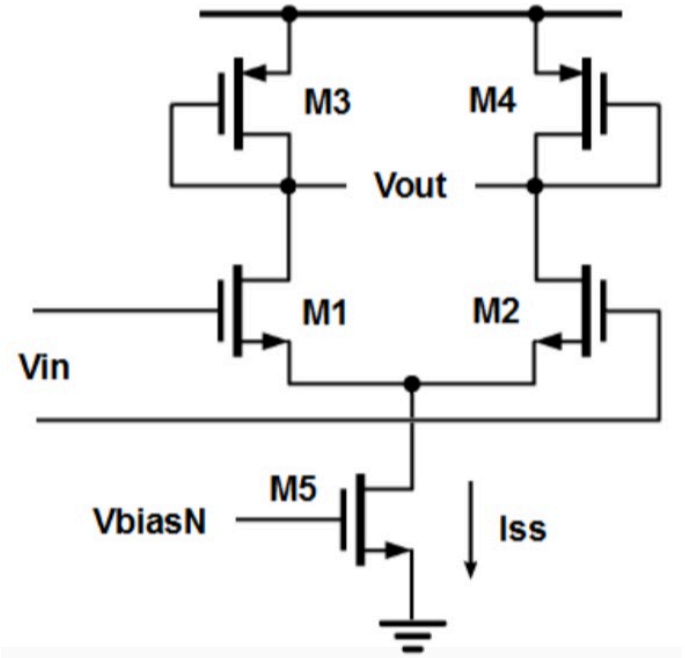

Figure 5.18: Schematic of the Differential Ampllifier without current splitting

In this configuration the small signal gain is still given by:

$$| A_0 | \simeq \frac{g_{m12}}{g_{m34}} \tag{5.4.1}$$

Nevertheless, the current flowing in M1-M2 and M3-M4 is now the same and the discrepancy between the transconductances is reduced. As a consequence, in order to recover some gain a different sizing of the transistors M3-M4 has been adopted, moving them from a $W/L = 0.3\ \mu m/1.5\ \mu m$ to a $W/L = 0.3\ \mu m/3\ \mu m$. In addition, the NMOS input pair has been cascoded, since this structure has a beneficial role in gain figures. The final implementation of the DA is presented in figure 5.19. The overall gain of this DA version is around 30% smaller with respect to the original design. Nevertheless, it allows to get almost rid of the gain variations observed in the first prototype. In fact, as figure 5.20 shows, the effective threshold dispersion diminishes drastically, with a RMS equal to 0.4 mV. It corresponds to around 5 electrons, which becomes a negligible contribution with respect to the ENC and the latch residual dynamic offset.

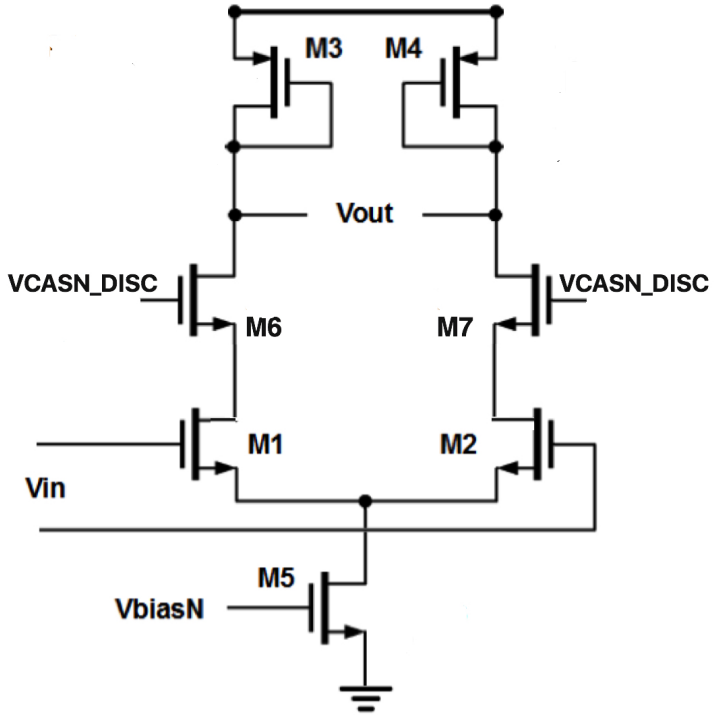

Figure 5.19: Schematic of the cascoded Differential Ampllifier

Device	Size (W/L) [$\mu m/\mu m$]
M1,M2	4/0.12
M3,M4	0.2/0.06
M5,M6	1.6/0.24
M7,M8	0.39/0.3
M9,M10	0.52/0.06
M11,M12	2/0.3

Table 5.1: Transistor sizing of the updated version of the positive feedback latch

5.4.2 Positive feedback latch

From the latch point of view, the simulations have shown that some devices have a strong impact of the dynamic offset, due to mismatch. In order to improve this quantity, it has then been necessary to increase their size. The updated dimensions of the positive feedback latch transistors are listed in table 5.1.
Therefore, in particular the NMOS and PMOS couples included in the positive feedback structures have been significantly enlarged. As figure 5.21 shows, these choices have allowed a substantial reduction of the dynamic offset with respect to the first prototype solution.

5.4.3 Layout

The layout of the analog front-end included in the CHIPIX_VFE2/TO prototype is presented in figure 5.22. The numbering of the building blocks is the same used for the first prototype. In addition to the modifications required by the new designs of DA and positive feedback latch, some additional optimizations have been performed. Each block has been designed with a rectangular shape, which allowed to realize an improved power and ground grid. In general, post-layout simulations have shown that the new layout has allowed a reduction of the parasitic

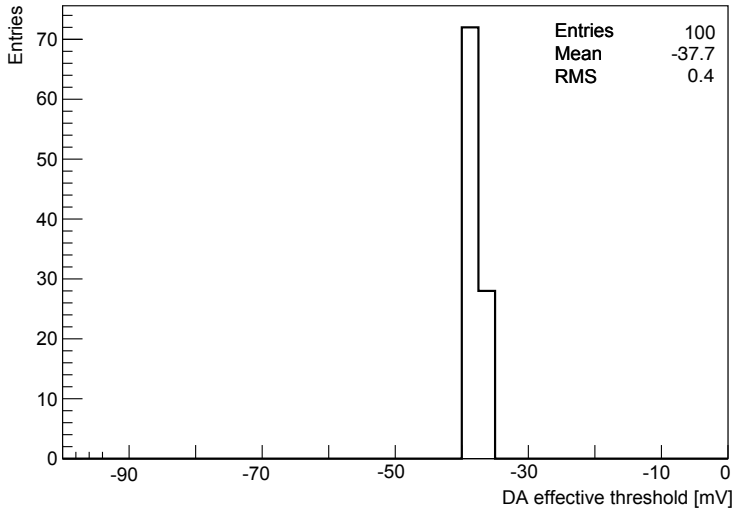

Figure 5.20: Monte Carlo simulations of the DA effective threshold for 100 pixels - Second prototype

elements in the preamp. As explained before, the AC coupling part has been modified: the resistor used in the first prototype has been removed and another capacitor of the same size has been connected to the second input of the DA in order to achieve a better symmetry. In addition, the positive feedback latch and the delay line have been included in a separate DNW, polarized with digital power, in order to provide a better isolation from the analog point of view.

5.4.4 Measurements

Given that the pad ring and therefore the output interface of the prototype has remained unchanged, the PCB already described was still compatible with the new chip. As a result, the same test setup has been used. Small improvements in the order of few percent have been spotted in the preamplifier parameters, thanks to the layout optimization. Therefore, after a first verification of the correct operation of the chip, the measurements have been focused on the discriminator, the only block containing major modifications with respect to CHIPIX_VFE1/TO. Using the S-curve technique, the threshold distribution for 16 pixels has been evaluated. As figure 5.23 shows, the modified DA and positive feedback latch have produced the expected beneficial effect. In fact, the threshold dispersion is reduced from 174 to 70 electrons. This value is compliant with the specifications. In fact, assuming an ENC equal to 90 electrons it is possible to write:

$$\sqrt{ENC^2 + \sigma_{thr}^2} = \sqrt{90^2 + 70^2} \simeq 114e^- \tag{5.4.2}$$

which is lower than the requirements of 126 electrons maximum specified inside the RD53 collaboration. In addition, the ENC can be further reduced by an accurate tuning of the Krummenacher current, which can be slightly reduced while keeping the dead time below 1%.

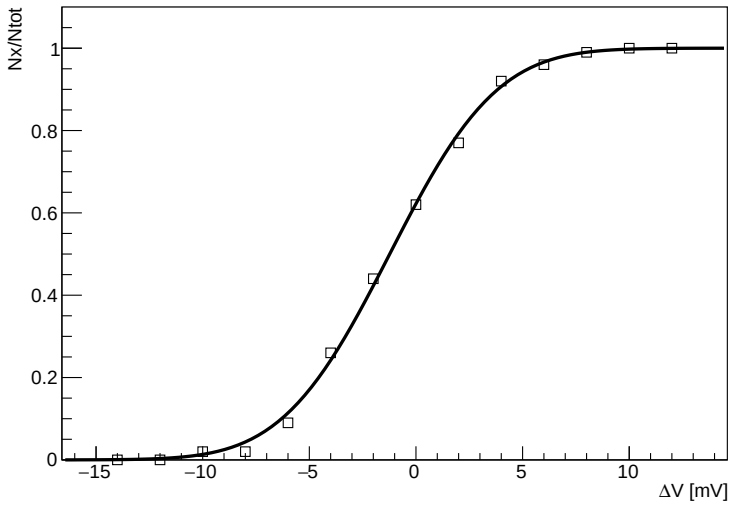

Figure 5.21: Representation of the latch dynamic offset in the second prototype

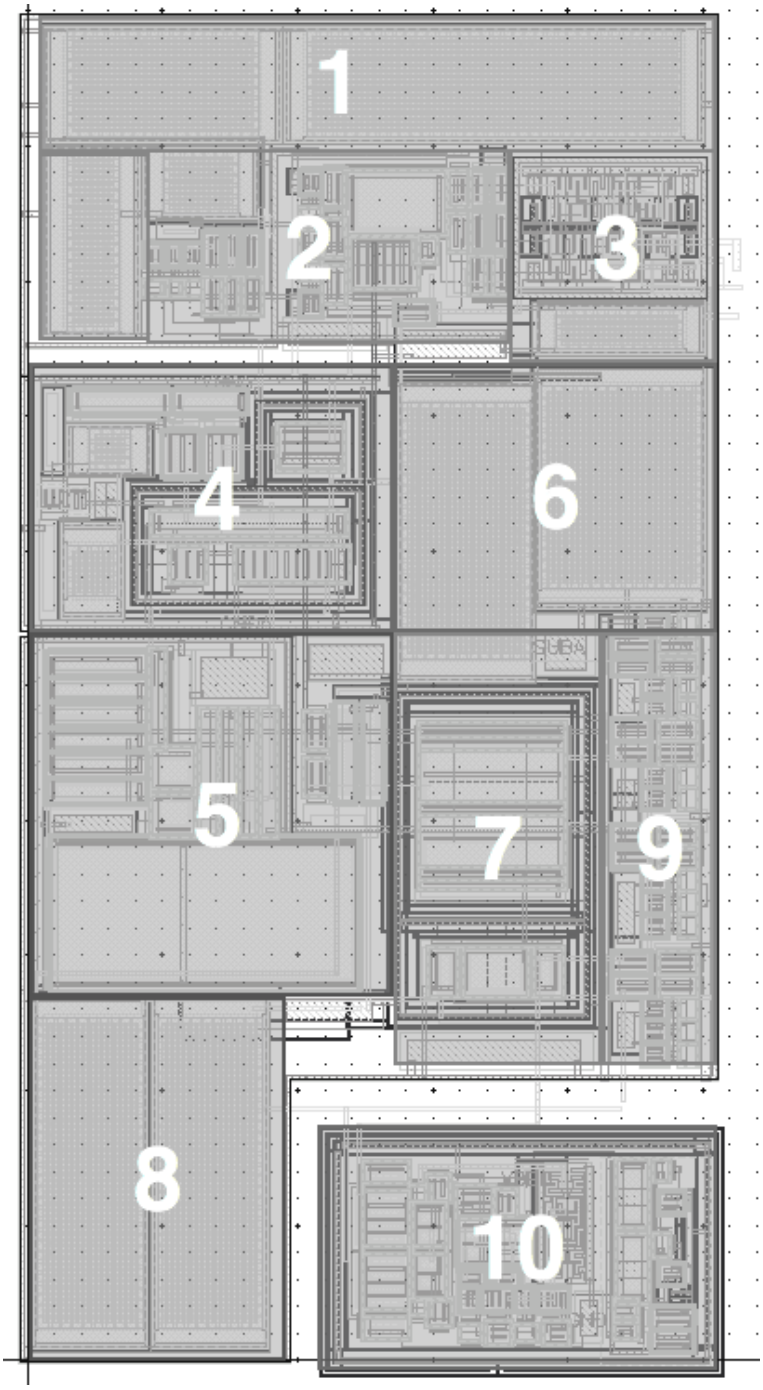

Figure 5.22: Layout of analog front-end included in the CHIPIX_VFE2/TO prototype

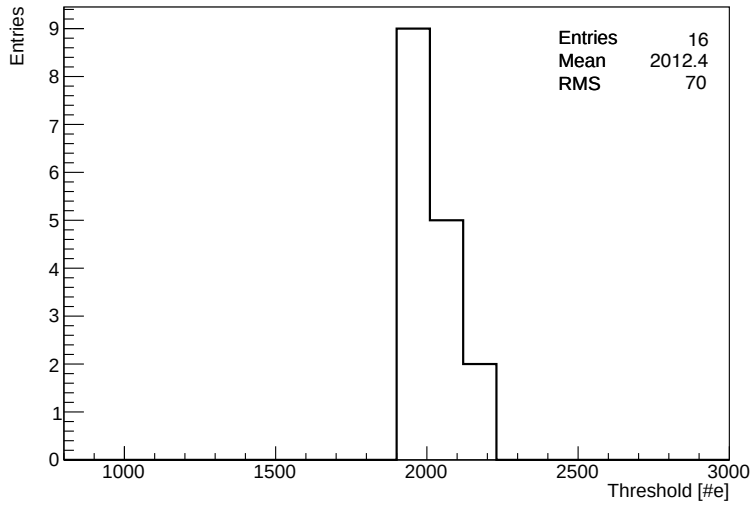

Figure 5.23: Threshold distribution for 16 pixels with offset compensation - Second prototype

5.5 Irradiation campaign

The CHIPIX_VFE2/TO prototype has been also tested under irradiation. The irradiation procedure has been performed at CERN, using a X-ray machine, reaching a TID equal to 600 Mrad. This level has been chosen keeping in mind the specifications about radiation tolerance for the readout chip at HL-LHC, which is at least 500 Mrad. A picture of the setup is provided in figure 5.24. The test board has been fixed to the support and then the X-ray source has been positioned so that a uniform irradiation is provided on the whole chip. This aspect is crucial since significantly different TID levels across the matrix can lead to an increased mismatch between pixels and therefore to misleading results. Given the small size of the prototype, it has been possible to move the X-ray source very close to the chip, at around 1 cm distance. In this way it has been possible to take advantage of the maximum radiation level provided by the machine, 9 Mrad/hr. The tests have been carried out at room temperature, since a cooling system was not available at that time. As a consequence, the obtained results represent a worst case compared to the normal operation inside the experiment: the low temperature, around -20 $°C$, is in fact expected to reduce the radiation effects. In addition, the chip has been kept in working condition during the whole irradiation phase, given that under bias the damage is expected to be higher.

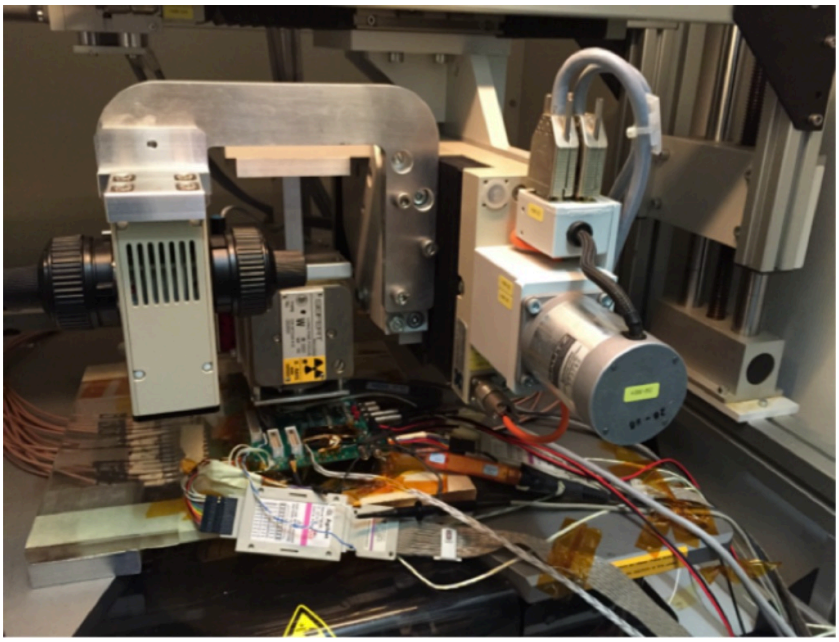

Figure 5.24: X-ray machine used for the 600 Mrad irradiation

Some quick measurements have been taken at fixed steps of TID in order to keep track of the evolution of the main analog parameters of the front-end with radiation. During these phases the X-ray machine has been paused in order to have the same TID across the measurements. Therefore, the In particular, after a first measurement before irradiation, a set of data taking has been carried out at 10, 20, 50, 100, 200, 390 and 600 Mrad.

5.5.1 Preamp

The amplitude of the preamplifier signal has been monitored by injecting a 5 ke^- input signal. Figure 5.25 shows that this parameter is barely influenced by radiation and remains almost constant through the whole process. Also annealing leaves the amplitude value essentially unchanged.

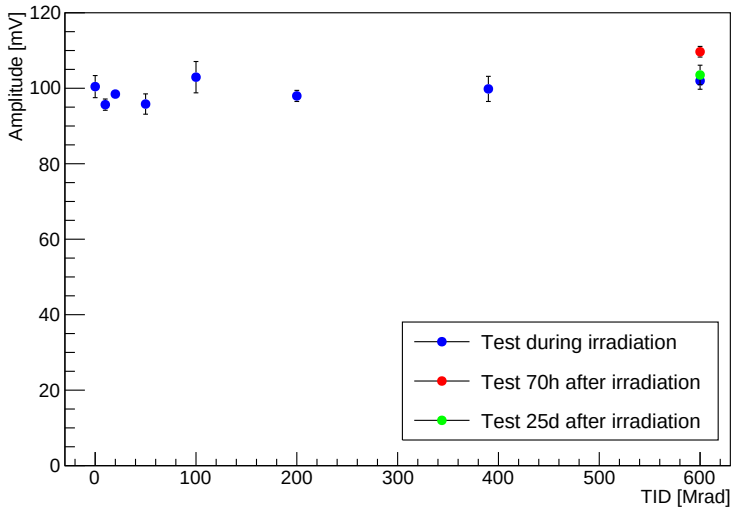

Figure 5.25: Preamplifier output signal amplitude as a function of TID

Significant variations have been instead spotted regarding the peaking time of the preamplifier output signal, as illustrated in figure 5.26. While keeping in mind that this value is measured at the output of the analog buffers and it therefore characterized by an increase in absolute value around 30%, the fluctuations of it with radiation give a hint of the actual increase of the peaking time in the pixel. Below 100 Mrad the peaking time value remains essentially constant. Then it starts to increase, with the most important variations taking place in the last step, between 390 and 600 Mrad. The 25-days annealing at room temperature provides a significant recovery, bringing back the peaking time close to the value at 390 Mrad.

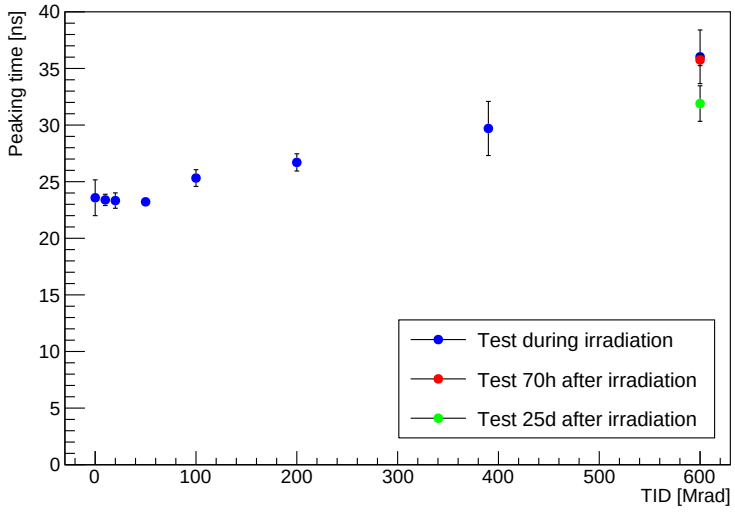

Figure 5.26: Preamp output signal peaking time as a function of TID

5.5.2 Discriminator

Figure 5.27 shows the ENC as a function of the input capacitance for a single pixel in three different configurations: before irradiation, immediately after the 600 Mrad irradiation and

after the 25-days annealing. While it confirms that the linear trend of the ENC is preserved in all cases, the plot highlights the significant increase in slope after irradiation. Although the increase in noise looks quite dramatic, it has to be underlined that with the expected values of the sensor capacitance, around 50 fF, the discrepancy between the two cases is limited to 10-15%. Annealing has a strong impact, moving the slope back to the original trend, but with an offset. As a result, around 50 fF it gives rise to almost no difference in comparison to the measurement taken immediately after irradiation.
From the point of view of threshold dispersion, a systematic measurement on the 16 pixels has not been performed due to the long time required by the complete S-curve data taking. Nevertheless, the point at efficiency 50% has been monitored for some pixels. It has shown a degradation in terms of the frequency of the offset compensation period. In practice, the discharge of the capacitors is accelerated by some source-drain leakage paths through the switches. For this reason, the implementation of these devices as ELTs has been envisaged for the CHIPIX65 and RD53A demonstrators presented in chapter 6.

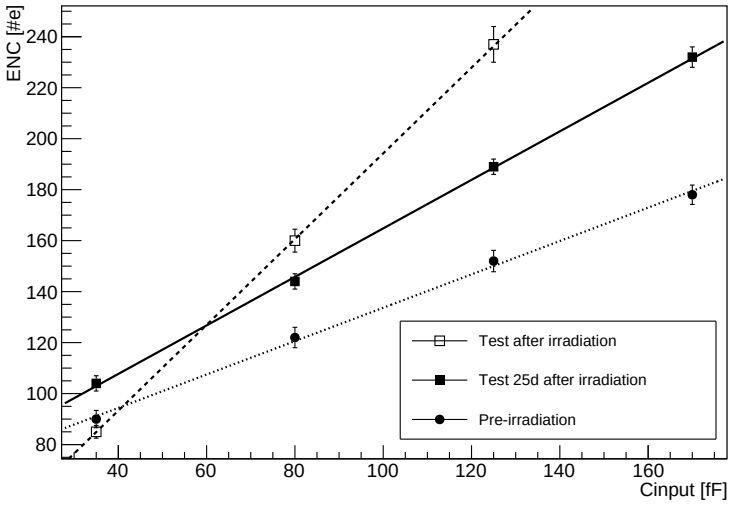

Figure 5.27: Comparison of the ENC vs input capacitance trends before irradiation, after irradiation and after annealing for a single pixel

Since the chip has been kept with the fast ToT option active during the whole irradiation campaign, the value of the oscillation frequency has been monitored at the oscilloscope. The trend of this parameter is presented in figure 5.28. The initial frequency before irradiation has been set at 140 MHz. At the beginning, between 10 and 20 Mrad, a small bump is observed and the frequency rises at almost 150 MHz. Then, starting from 50 Mrad, a roughly constant decrease is observed, reaching 110 MHz at 600 Mrad. Similarly to the previous parameters, annealing has a beneficial effect and the frequency rebounds to more than 120 Mrad after 25 days at room temperature.
This trend is explained by the effects seen of single transistor studies presented in chapter 2. In fact, in the delay line stage, which is responsible of the frequency tuning, quite small devices are used. At low irradiation levels, around few Mrads, radiation-induced short channel effects play an important role and the NMOS transistors tend to drive more current [73]. As a consequence, an increase of the current in the delay line results in an increase of the oscillation frequency. At higher radiation doses, instead, the situation reverses due to the interface states and the current decreases almost linearly. A further confirmation of this trend comes from a digital test prototype designed inside the RD53 collaboration for inspection of the radiation damage effects on 65 nm digital circuitry, the DRAD chip [74]. It has been irradiated in different campaigns

at both room and cold temperature (-20 °C) and up to 500 Mrad. Almost all the standard digital cells exhibit an improvement in time delay at few Mrads before experiencing a constant degradation starting from values around 20 Mrads, showing therefore a behavior compatible with the one presented in figure 5.28.
It has nevertheless to be recalled that the current flowing in the delay line implemented in the front-end can be tuned by adjusting the settings in the bias cell. In fact, since the frequency distribution remains as good as before irradiation, a global optimization of the current is propagated to all the pixels with small fluctuations, allowing to compensate the variations induced by radiation.

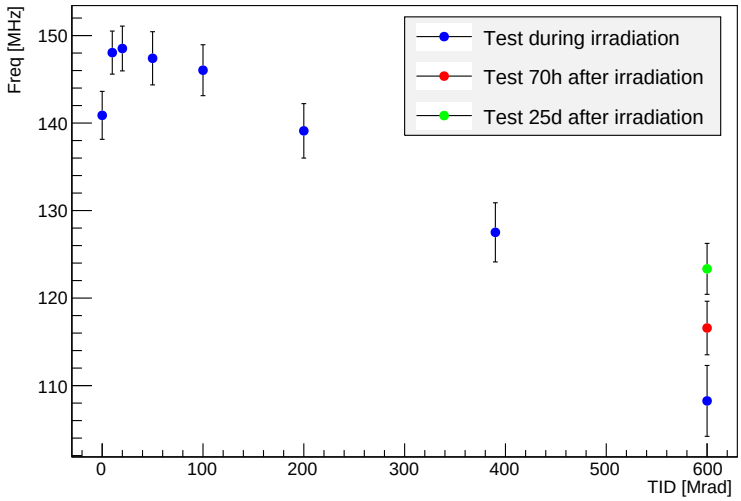

Figure 5.28: ToT frequency as a function of the TID

5.6 Summary

The performance of the CHIPIX_VFE2/TO prototype has represented a clear improvement with respect to the CHIPIX_VFE1/TO. It has allowed in fact to keep the good figures in terms of gain, noise and ToT together with a substantial reduction of the threshold dispersion. Also the irradiation measurement have shown good results, with acceptable degradation in some parameters like the preamplifier signal peaking time and noise. The most critical point appears to be the offset compensation capacitor discharge due to leakage currents in the switches. The use of ELTs, as shown in chapter 6, reduces the degradation. As a consequence, the version of the analog front-end included in this chip has been taken as the baseline for the CHIPIX65 and RD53A demonstrator chips described in the next chapter.

Chapter 6

CHIPIX65 and RD53A demonstrator chips

In this chapter two prototypes, the CHIPIX65 and RD53A demonstrator chips are presented. They feature a structure closer to the final chip implementation, with a full digital architecture in both the pixel matrix and the periphery, together with a programmable biasing and monitoring scheme. The last year of the Ph.D. activity has been focused on the integration of the synchronous analog front-end presented in this work inside these prototypes.

6.1 The CHIPIX65 demonstrator

In view of the RD53 collaboration full scale chip, the INFN CHIPIX65 project has designed a $3.5 \times 5.1\text{mm}^2$ prototype, called CHIPIX65 demonstrator [75]. Its main purpose is to test all the building blocks built by the INFN groups in a chip having a level of complexity comparable to a final design, in view of the RD53A large scale prototype. The layout of the demonstrator is presented in figure 6.1. It is characterized by a 64×64 pixel matrix which contains two flavors of analog front-ends and a digital architecture organized in 4×4 pixel regions. Along with the analog front-end presented in this work, an asynchronous architecture, which is described in detail in [76], has been included. Each of the two analog designs occupy a half of the matrix. The chip periphery contains the blocks needed for the biasing scheme. Each current and voltage required by the front-end is controlled by a 10-bit DAC [77] which can be programmed for outside the chip. The outputs of the DAC are then fed into a network of current mirrors, designed in order to provide the different currents with the requested resolution and to guarantee linearity in the full operating range. These mirrors are connected to the column bias cells, designed to drive 64 pixels. During operation the DAC needs a reference current, provided by a bandgap reference circuit (BGR) [78]. In addition, a 12-bit ADC has been integrated on-chip for monitoring DC voltage levels and slow varying signals. The chip periphery contains also the end of column digital logic required for the readout.

Inside the matrix, the basic building block is the 4×4 pixel region, since it contains the single unit from the point of view of the digital architecture. It is then replicated many times to obtain the full pixel matrix. From the point of view of the matrix columns, the "analog island" approach represented in figure 6.2 has been adopted, given that it will be used also for the RD53A chip. It consists in organizing the area in the single pixel so that the analog front-end is contained in a $35 \times 35\ \mu m^2$ square. The structure is then mirrored in the other directions to form an analog area surrounded by digital cells. In addition, the whole analog part is included in a DNW and surrounded by a guard-ring. The idea is therefore to provide the best isolation from substrate noise.

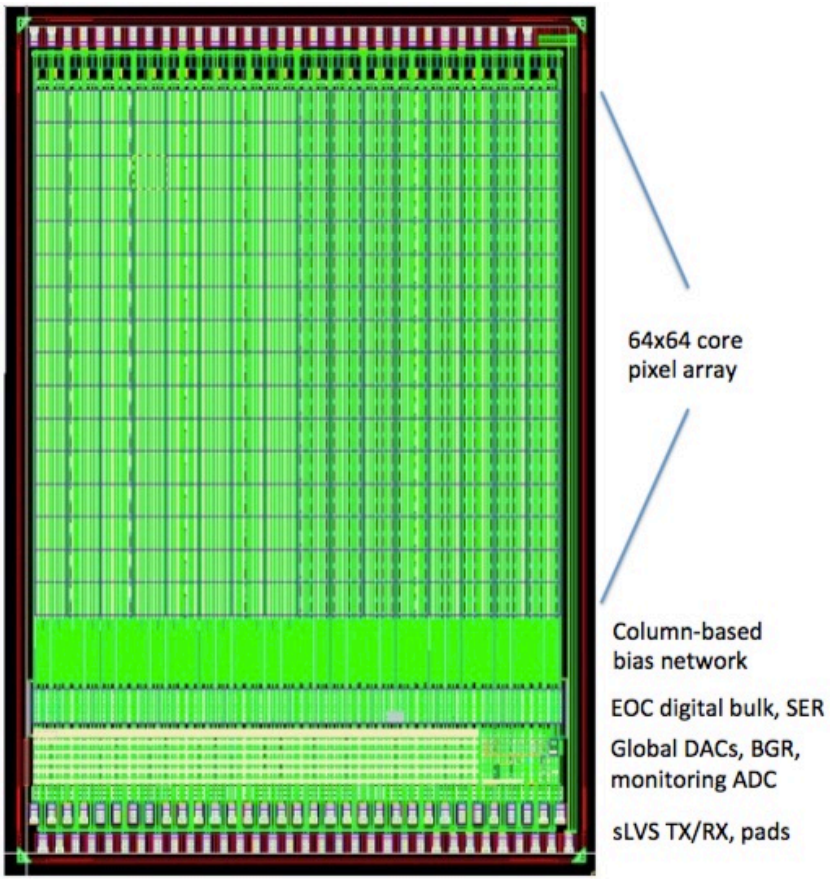

Figure 6.1: Overview of the CHIPIX65 demonstrator

6.1.1 Optimization of the synchronous analog front-end

Based on the results obtained with the CHIPIX_VFE1/TO and CHIPIX_VFE2/TO test prototypes, the design of the synchronous analog front-end has been further optimized. Firstly, all the CMOS switches involved in the offset compensation phase have been designed as Enclosed Layout Transistors. In fact, as demonstrated by the radiation tests on single transistors illustrated in chapter 2, the standard NMOS transistors exhibit significant variations of the source-drain leakage current, with little dependence on the device width. ELTs show instead an almost constant trend. This modification should then minimize the discrepancy in the capacitors discharge time between the non-irradiated and the 500 Mrad configurations.
Another key aspect is the pixel layout. In fact, in order to be compatible with the $35 \times 35\ \mu m^2$ configuration, the placement of the building blocks has been modified. Since the CHIPIX65 demonstrator will be bump-bonded to a sensor after a first testing phase, the local input capacitors have been removed. In addition, also the analog buffer is not included, given that only the discriminator output is studied. These choices have also allowed to save a significant percentage of the area. Therefore, the insertion of the analog front-end in the prescribed square did not require any additional modification.

6.1.2 CHIPIX65 demonstrator test setup

A custom test PCB has been designed for the CHIPIX65 demonstrator tests. In comparison with the one used for the small prototypes, this board is simpler because the bias and configuration section inside the chips needs only a pattern of digital signals which are provided by a FPGA to which the board is connected. The main purpose of the PCB is therefore to provide

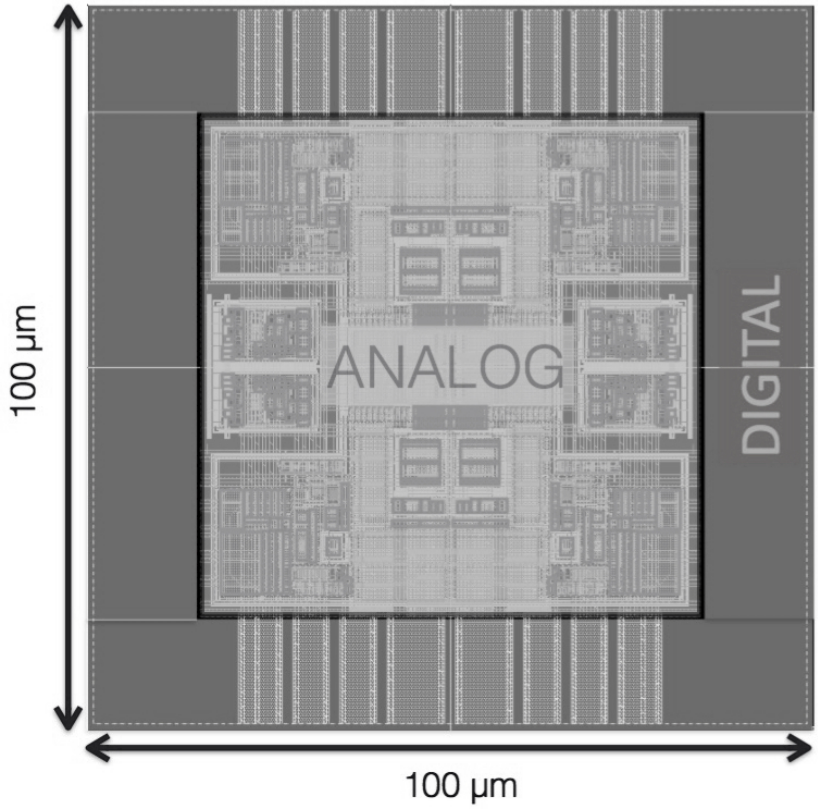

Figure 6.2: Layout of a 2×2 pixel region

the separate analog and digital power (and ground) and some test points for monitoring, together with the required connections between the chip and the FPGA. A Virtex-7 Xilinx board has been used for the configuration. A picture of this setup is given in figure 6.3. For the data acquisition, a Labview-based setup has been used. Using this interface, it has been possible to control all the currents and voltages, together with the other functionalities like gain and fast oscillation. The main window of the program is presented in figure 6.4. The interface allows for example to produce the S-curves described in the previous chapter by performing a parametric sweep of the threshold voltage or of the input signal. It is also possible to modify two parameters simultaneously, obtaining the S-curves family illustrated in the figure. The data are then stored in files that can be analyzed offline. In this way a very efficient automatic data acquisition is possible, allowing to improve drastically the statistics with respect to the small test prototypes.

6.1.3 Test results

The first tests have shown that all the pixels are fully working. The correct functionality of all the building blocks included in the demonstrator has been verified. In particular, a characterization of the linearity of the DACs has been performed in order to check the values of the currents and voltages, obtaining results compliant with the simulations. At this point a proper characterization of the pixel matrix has been possible.

Threshold distribution

At this point a systematic characterization using the S-curve technique has been performed by fixing the threshold DAC value and scanning the input charge. All of them are taken with the high value of the Krummenacher current for the fast ToT. A superimposition of the results

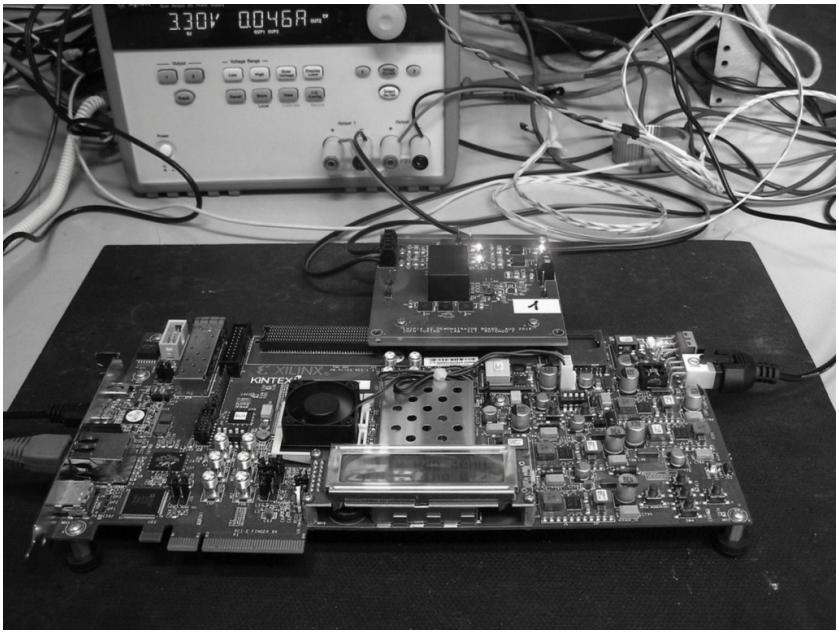

Figure 6.3: Test setup for the CHIPIX65 demonstrator

obtained for all the pixels is shown in figure 6.5. The same procedure has been repeated for different thresholds, indicated with both their values in DAC counts and electrons, which is the most relevant quantity. The resulting distributions are reproduced in figure 6.6. It shows that the pixels can be operated at very low thresholds (around 300 electrons) without losing in performance. The threshold dispersion is around 100 electrons, which is a larger value with respect to the measurement on the CHIPIX_VFE_2/TO prototype. Nevertheless, it has to be underlined that the statistics was quite limited, while in this case 1024 pixels have been evaluated. The mean values of these distributions have been then used to produce the plot shown in figure 6.7. It provides the relationship between the threshold voltage and the input charge which has been necessary in order to express the threshold value in electrons. It shows that a linear trend, as expected, is obtained.

Noise

The same family of S-curves has been used to draw some considerations about noise. As figure 6.8 shows, the variation of the threshold voltage has little influence on noise. In fact, the four distributions are almost superimposed. They also indicate that the noise variations between pixels are under control, given that the RMS of the distribution is equal to few electrons. This aspect is confirmed by the plot of the ENC as a function of the threshold value illustrated in figure 6.9. In fact, no specific trend is obtained. The resulting ENC values, all between 80 and 100 electrons, are compatible with the expectations.

ToT

From the point of view of the ToT, a comprehensive study has been performed by enabling the fast oscillator. The digital architecture contained in the pixel allows the digitization of this information by using a 5-bit counter. The values of the latter has been then used for the characterization. Figure 6.10 shows the ToT value as a function of the input charge. The expected linear trend is still verified. The same procedure has been repeated for all the pixels, giving rise to the plot presented in figure 6.11. The superimposition of all the characteristics shows that the spread of the slope between the pixels is significative. In fact, as indicated by

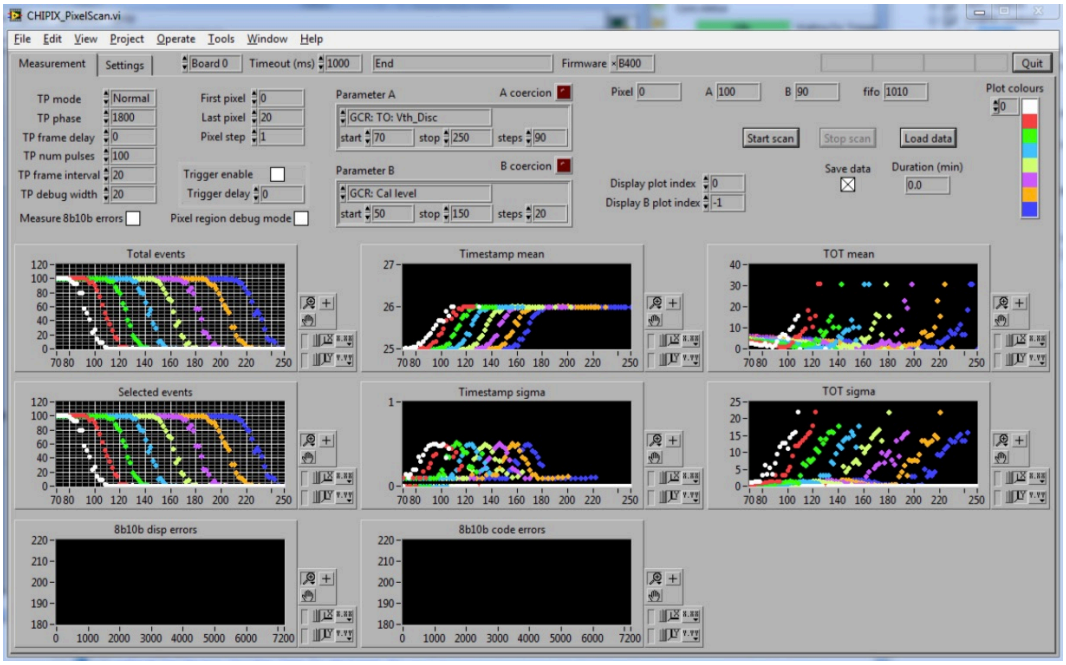

Figure 6.4: Data acquisition interface

the histogram illustrated in figure 6.12, the RMS of the distribution is equal to around 13% of the mean value. Nevertheless, this situation was expected from simulations: it is caused by the mismatch in the Krummenacher feedback. It can be reduced only by increasing significantly the transistor sizes, but such a modification is not compatible with the area available. These results are also compatible with the results obtained by the other front-end included in the demonstrator. In addition, this fluctuations can be mitigated by performing offline corrections based on the characterization of the chip. As a consequence, these results can be considered acceptable.

Summary

In summary, the first characterization of the analog front-end inside the CHIPIX65 demonstrator have shown very promising results, compliant with the expectations. The main features of the front-end, like the offset compensation and the fast oscillator, work properly in all the pixels. The average ENC is 90 electrons, with a threshold dispersion around 100 electrons. For the former, it can be further decreased by setting the ideal trade-off in the Krummenacher current between dead time and noise. For the latter, an additional optimization of the latch sizing is envisaged for the RD53A prototype. Nevertheless, it is already a result consistent with the specifications. In addition, when the CHIPIX65 demonstrator will be bump-bonded to a sensor, an extensive characterization will be performed, featuring also test beams.

6.1.4 Irradiation campaign at -20 °C

An irradiation campaign has been performed using the CERN X-ray machine, this time at cold temperature (-20 °C). The final TID is 600 Mrad, the same as for the CHIPIX_VFE2/TO prototype. During the whole irradiation period the chip has been kept under working conditions, with continuous data taking. Measurements have shown that the chip functionality is completely preserved at 600 Mrad, on both analog and digital sides. Some characterizations have been then performed in order to inspect the degradation induced by radiation on some key parameters like noise and threshold dispersion.

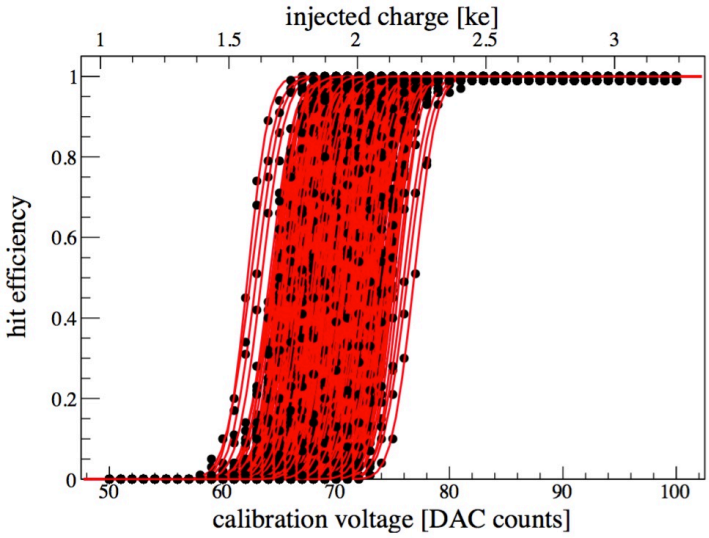

Figure 6.5: S-curves for 1024 pixels

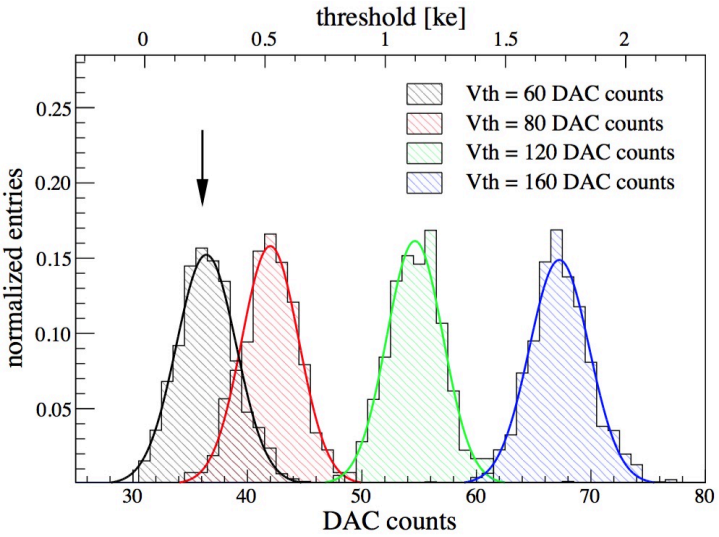

Figure 6.6: Threshold dispersion for different values of the threshold DAC

A first set of plots has been carried out by superimposing the results taken at room temperature before irradiation and after the end of irradiation. Figure 6.13 shows the trend of the threshold dispersion for different values of the global threshold. An increase of this quantity in the order of 20 electrons is obtained for all the points. Therefore, the use of ELTs allows to maintain the threshold dispersion under control. It has also to be underlined that these results are taken immediately after the end of irradiation. As a consequence, no annealing has been carried out. The latter is expected in fact to reduce the discrepancy compared to the pre-irradiation case.
The same S-curves have been used to measure the noise. The results are illustrated in figure 6.14. At low threshold voltages, basically no variation in the value of the ENC is obtained. On the other hand, for the two higher thresholds the ENC is characterized by a 10% increase. Also in this case, based on the irradiation of the CHIPIX_VFE_2/TO prototype, an improvement with annealing at room temperature is expected.
Regarding the measurements taken at -20 °C, some differences have been obtained. From the point of view of the threshold dispersion, as shown in figure 6.15, at very low threshold the increase with radiation is very limited, while it is more remarkable with higher DAC values. On

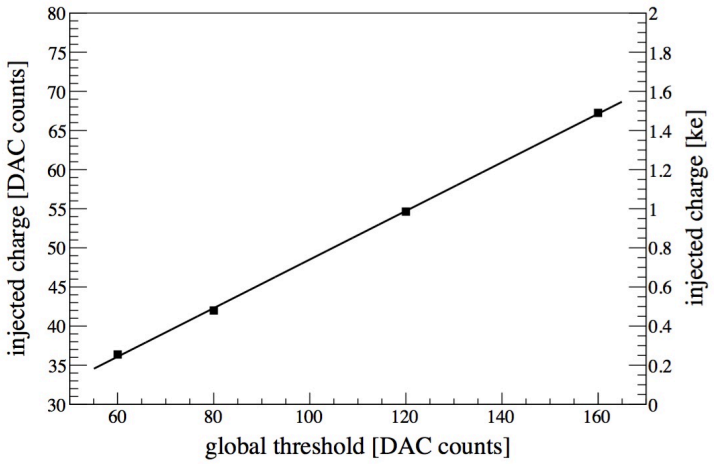

Figure 6.7: Threshold versus input charge characteristic

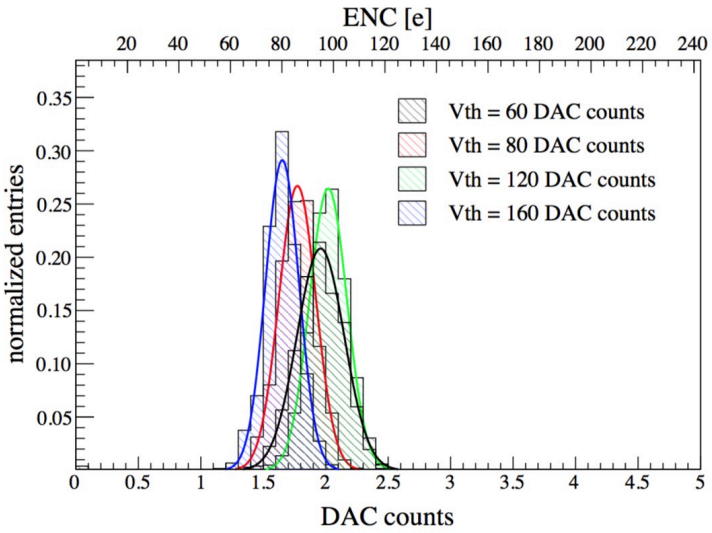

Figure 6.8: Noise distribution for different thresholds

the other hand, the ENC is characterized by an around 10% increase in all the points expected the lowest one, in which is limited, as shown in figure 6.16. Nevertheless a reduction of noise in absolute value, in the order of 10 electrons, is obtained thanks to the low temperature operation.
The next step is then to repeat the same set of measurements after a one month annealing at room temperature, in order to check if the expected recovery takes place. These first measurements have however shown that the chip can still operate after 600 Mrad, which is higher than the minimum specification for RD53, with a degradation of its main parameters below 20% at the end of the irradiation campaign.

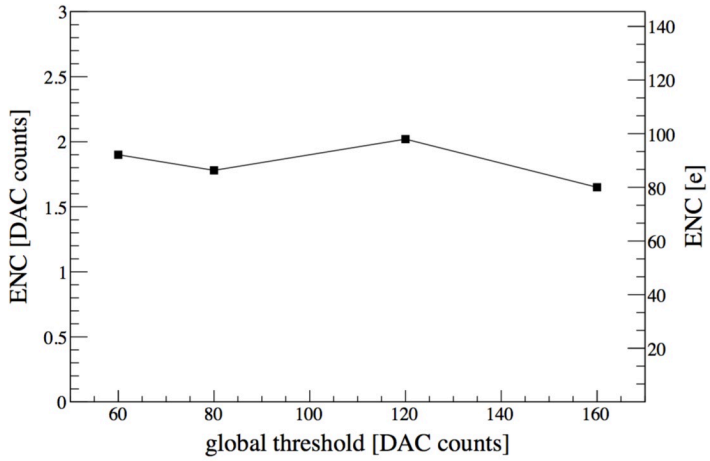

Figure 6.9: ENC as a function of the threshold value

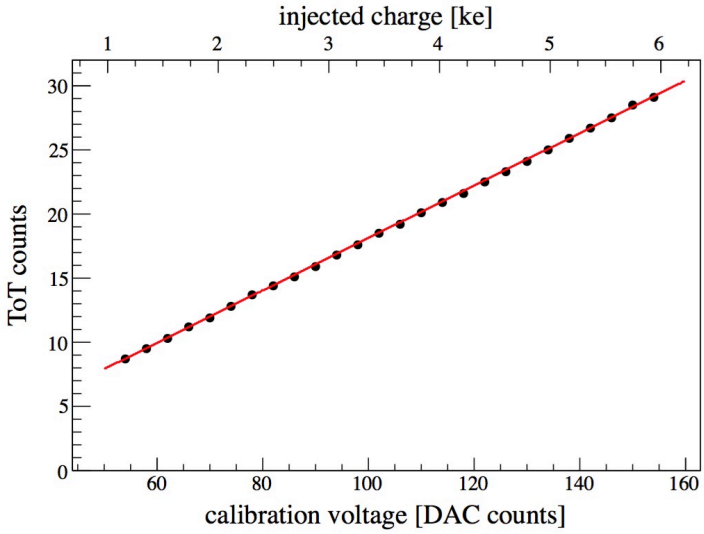

Figure 6.10: Data acquisition interface

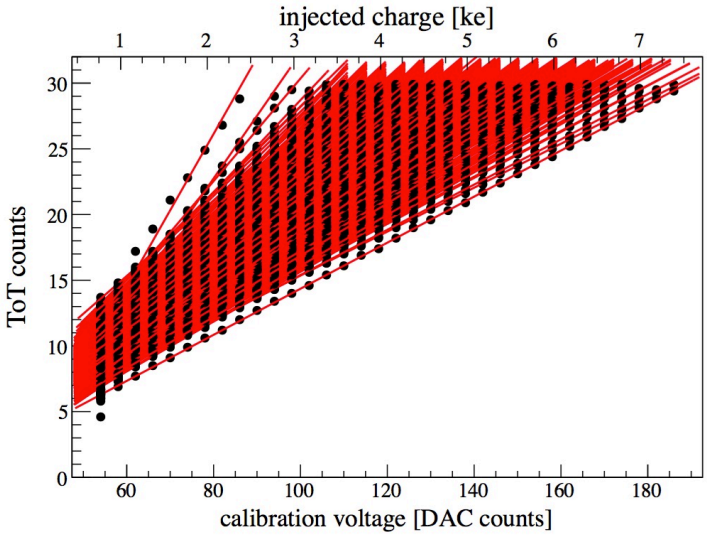

Figure 6.11: Data acquisition interface

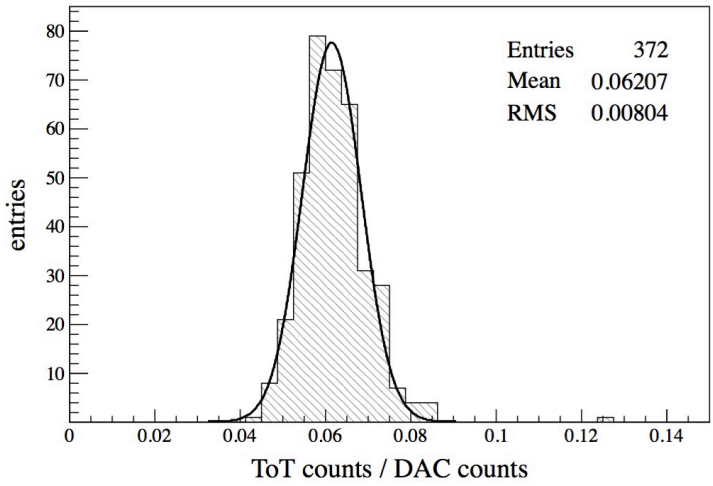

Figure 6.12: Data acquisition interface

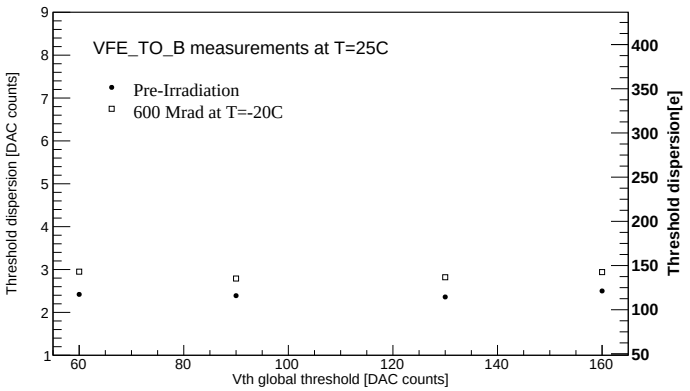

Figure 6.13: Threshold dispersion at room temperature before and after the end of irradiation

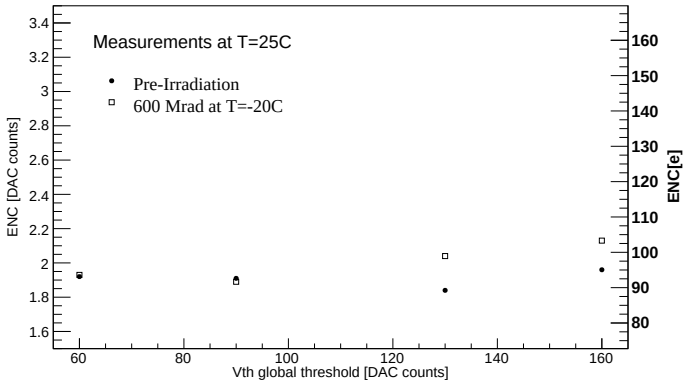

Figure 6.14: ENC value at room temperature before and after the end of irradiation

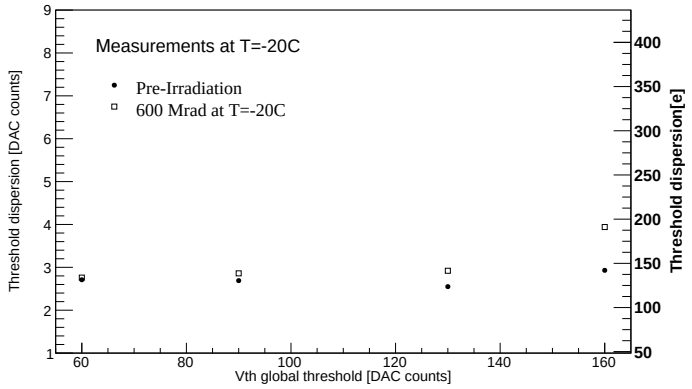

Figure 6.15: Threshold dispersion at -20°C before and after the end of irradiation

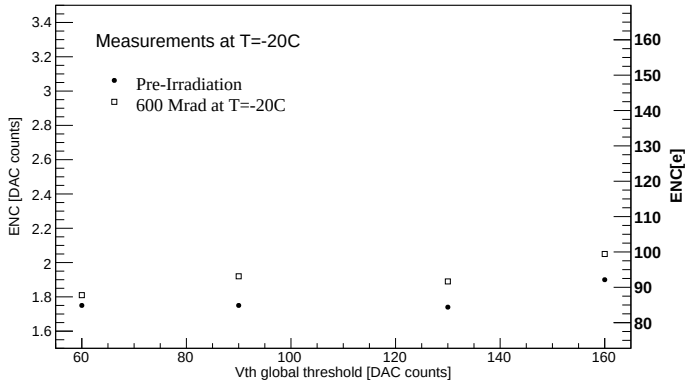

Figure 6.16: ENC value at -20°C before and after the end of irradiation

6.2 The RD53A demonstrator

The RD53A prototype, having a size of $20 \times 11.8\ mm^2$, is undergoing the final design phase. The main goal of this prototype is to demonstrate the suitability of the chosen CMOS 65 nm process for the Phase 2 silicon pixel detectors of CMS and ATLAS. In addition, it will allow to perform a full characterization of he building blocks implemented inside the RD53 collaboration. Therefore, the chip is being designed to provide extensive testing capabilities. With respect to the CHIPIX65 demonstrator, conceived as an intermediate step towards RD53A, the latter is significantly larger. It in fact contains 192×400 pixels in the core matrix. In addition to the two analog front-ends being part of the CHIPIX65 chip, a third design has been included. It is an asynchronous architecture already submitted and tested inside the FE65-P2 prototype. The latter is a small demonstrator having the same size with respect to the CHIPIX65 design. It is an evolution of the FE-I4 [79] already introduced in the present ATLAS Insertable B-Layer (IBL), with a strong contribution of the institutes which designed it and are now part of the RD53 collaboration [80]. FE65-P2 contains also a digital architecture based on a 2×2 pixel region scheme with a distributed latency buffer, which will be part of RD53A along with the one included in CHIPIX65. In RD53A it has also been decided to take already into account the effects that will be present in the final chip, which will feature a 400×400 pixel matrix, in order to perform the least possible modifications between the two designs.

6.2.1 Modifications in the synchronous front-end

A first aspect to the considered for the RD53A prototype is the voltage drop on the analog power and ground rails. In fact, the columns are significantly longer than in the CHIPIX65 demonstrator and in the final chip, with 400 pixels, this phenomenon will be even stronger. In order to understand in detail this effect, it has to be considered that the power grid supplies a double column. Therefore, it is possible to assume, from an analog point of view, the 4-pixel analog island as the fundamental cell. The voltage drop along the column is then given by the following relationship:

$$V_{drop} = \frac{R_{quad}\ I_{quad}\ n(n+1)}{2} \tag{6.2.1}$$

in which I_{quad} is given by the current flowing in the four pixels:

$$I_{quad} = 4\ I_{pixel} = 16\ \mu A \tag{6.2.2}$$

R_{quad} is instead the resistance of the metals used in the portion of power grid present in the 2×2 pixel region considered:

$$R_{quad} = \rho_{metal} \frac{L_{quad}}{S_{metal}} \tag{6.2.3}$$

The metal choice in the power grid has been performed in order to minimize this quantity, leading to a $R_{quad} = 28\ m\Omega$. As a consequence, for a column featuring 96 analog islands the voltage drop is:

$$V_{drop,96} = \frac{28 m\Omega\ 16 \mu A\ (96 \times 97)}{2} = 2.09\ mV \tag{6.2.4}$$

Considering instead the full column, composed of 200 analog islands, it becomes:

$$V_{drop,200} = \frac{28 m\Omega\ 16 \mu A\ (200 \times 201)}{2} = 9\ mV \tag{6.2.5}$$

While the first one is quite small, the second is relevant. Since one of the RD53A goals is to prepare a design compatible with a full scale chip, this aspect has to be addressed. In fact, even voltage drops of few mV affect the performance of the front-end, inducing variations in the value of the bias currents. In particular, simulations show that the Krummenacher feedback

is the mostly affected cell, since it is characterized by a very low current, in the order of tens of nA. This block is sensitive to the drop on the ground rail. The latter it is responsible to a reduction of the Krummenacher current as the drop increases along the column. Therefore, it results in an increase of the ToT dispersion across the matrix. The impact of voltage drop on the Krummenacher feedback performance can be understood by performing a mismatch simulation across 100 pixels. In case no voltage drop is applied, the distribution of the ToT presented in figure 6.17 is obtained. If then the 2.09 mV voltage drop expected for RD53A is considered, it gives rise to the distribution illustrated in figure 6.18. It is therefore the histogram of 100 pixels at the end of the column, where the drop is maximum. The increase of the mean value, caused by the reduction of the current, is however very limited. The same applies for the sigma. Nevertheless, if the final chip configuration with 9 mV drop is considered, the increase of mean and sigma is relevant. As figure 6.18 shows, both values increase by almost 20% compared to the case without voltage drop. It leads therefore to an even larger dispersion across the matrix, given that the mean and sigma increase along the column. For this reason, it has been decided to provide a separate metal line dedicated only to the grounding of the Krummenacher feedback circuit. In this way, given that the Krummenacher current for the single pixel is limited (40 nA maximum), the voltage drop effect across this line can be kept below 1 mV. As a consequence, the distribution illustrated in figure 6.17 can be applied not only to the first row but to the whole matrix, avoiding an additional increase of the ToT dispersion.

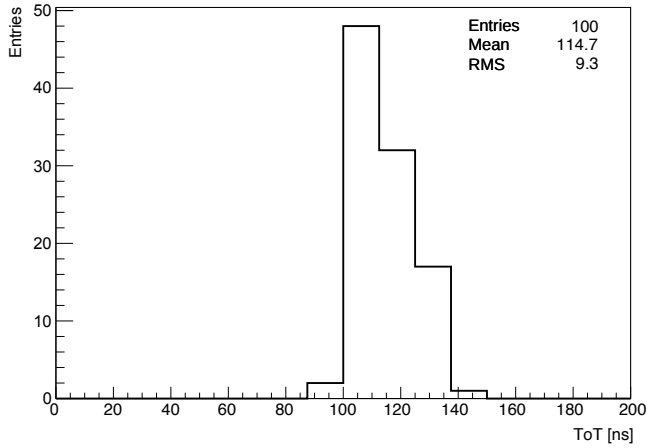

Figure 6.17: Distribution of the ToT for a 10 ke$^-$ input charge for 100 pixels without voltage drop

In addition, some improvements in the layout have been made, concerning the parasitic capacitances in the positive feedback latch. It has allowed to reach a 10% reduction of the power consumption of the stage.
After the submission of the chip, planned in Spring 2017, an extensive characterization campaign, including irradiation tests, will be performed. Subsequently, the chip will be bonded to dedicated sensors to perform full system measurement, together with test beams.

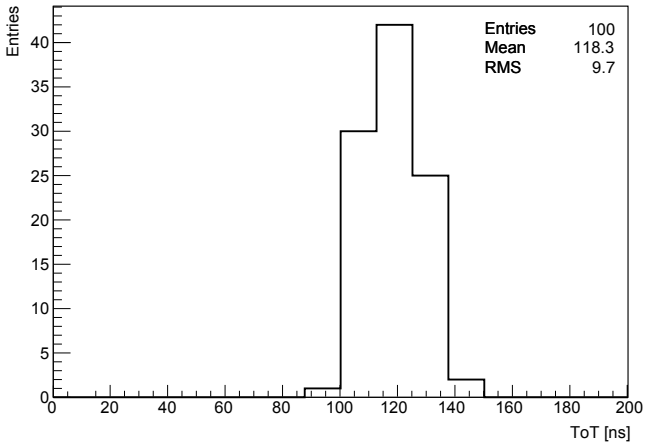

Figure 6.18: Distribution of the ToT for a 10 ke$^-$ input charge for 100 pixels with a 2 mV voltage drop

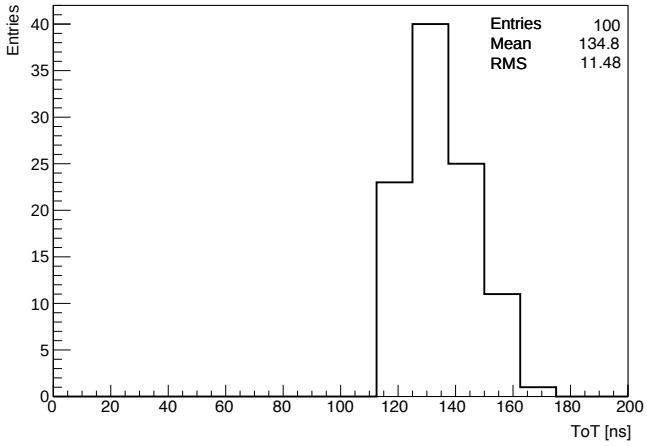

Figure 6.19: Distribution of the ToT for a 10 ke$^-$ input charge for 100 pixels with a 9 mV voltage drop

Chapter 7

Conclusions

In view of the Phase 2 upgrade of the CMS silicon pixel detector, an innovative front-end architecture has been developed in CMOS 65 nm technology. It is composed of a single stage Charge Sensitive Amplifier with the Krummenacher feedback scheme and a synchronous discriminator. The offset of the Differential Amplifier is compensated by hardware using switched capacitors. In this way the long trimming procedure required by the usually implemented local DACs is not necessary and the offset compensation can be performed during the abort gap of the machine, without increasing the dead time of the front-end. The proper discrimination is then performed by a positive feedback latch in correspondence of the rising edge of the 40 MHz clock. In addition, this stage can be turned into a local oscillator by means of an asynchronous logic feedback loop. This technique, often applied by high frequency applications like SAR-ADCs, allows to perform a fast ToT with tunable frequencies even beyond 500 MHz. Therefore, this architecture allows to measure the charge on a large dynamics (at least in the 1-40 ke^- range).
After a comprehensive characterization of the front-end through CAD simulations, an on-silicon measurement campaign has been performed. Firstly, two stand-alone small prototypes featuring a 8×8 pixel matrix have been submitted for production to the foundry. The first one, called CHIPIX_VFE_1/TO, has delivered very promising results. In fact, all the pixels were fully working, also regarding the offset compensation and the fast oscillator. From the preamplifier point of view, a satisfying gain distribution has been obtained, with a RMS equal to 2.2% of the mean value. Regarding the discriminator, the ENC linearity as a function of the input capacitance has been verified. For a typical sensor capacitance equal to 50 fF, the ENC is between 80 and 100 electrons depending on the Krummenacher current value. In addition, the ToT linearity has been verified in the whole range of interest of the input charge. All the test results show a very good agreement with the simulations. The second prototype, called CHIPIX_VFE_2/TO, has allowed to improve the threshold dispersion figure. Measurements have shown that it has been in fact reduced from the 174 electrons found with CHIPIX_VFE_1/TO to 70 electrons. CHIPIX_VFE_2/TO has also undergone a X-ray irradiation campaign up to 600 Mrad. It has led to a light degradation in the CSA peaking time, noise and ToT frequency, but always at levels compatible with the requirements. The offset compensation switches seem affected by leakage increase and therefore they have been implemented with ELT devices inside the CHIPIX65 demonstrator. The latter features the integration of the synchronous front-end with an innovative digital readout architecture. The bigger size of the matrix (64×64) has allowed to increase the number of pixels involved in the measurements. Nevertheless, the results are compatible with the ones obtained with the small prototypes. An irradiation campaign at low temperature up to 600 Mrad has confirmed a small degradation of the front-end main parameters, however aways below 20% with respect to the pre-irradiation case.
These results have therefore confirmed the suitability of the synchronous architecture for the

Phase 2 silicon pixel detector readout. Therefore, it is one of the solutions implemented in the RD53A demonstrator, which is going to be submitted to the foundry in Spring 2017. It has been designed taking already into account the effects expected in the final chip, like voltage drop on the power and ground lines. Therefore, the precaution of a separate ground line with reduced drop has been taken in the front-end on the most sensitive block, the Krummenacher feedback. Both the demonstrators will be also bonded to sensors in order to have a full-system characterization with test beams.

Acknowledgements

I want to express a special thanks to my Ph. D. supervisor, Angelo Rivetti, for the support and the advices that he has given me during the last three years. I would also like to thank Natale Demaria, Luca Pacher and Andrea Paternò for all the time spent together for the design, simulation and testing of the prototypes described in this work. Without this collective effort this work would not have been possible.
I want also to thank all the members of the Torino VLSI group and of the electronics laboratory, in particular Francesco Rotondo for the PCB design and Richard Wheadon for the LabView acquisition interface. A word of thanks then to the CMS group in Torino and the people involved in the CHIPIX65 and RD53 collaborations for the fruitful discussions and sharing of information.
A big thanks to all the Ph.D. students with whom I have spent the last three years. Finally I want also to thank professor Ezio Menichetti for having done an internal review of this work.

Bibliography

[1] cms.web.cern.ch.

[2] http://home.cern/topics/large-hadron-collider.

[3] The CMS collaboration, "CMS Physics Technical Design Report, Volume I: Detector Performance and Software," tech. rep., CERN, 2011.

[4] https://project-hl-lhc-industry.web.cern.ch/content/project-schedule.

[5] The CMS Collaboration, "The CMS experiment at the CERN LHC," *Journal of Instrumentation*, vol. 3, 2008. doi:10.1088/1748-0221/3/08/S08004.

[6] The CMS Collaboration, "CMS tracking performance results from early LHC operation," *European Physics Journal C*, vol. 70, 2010. doi:10.1140/epjc/s10052-010-1491-3.

[7] The CMS Collaboration, "Radiation experience with the CMS pixel detector," *Journal of Instrumentation*, vol. 10, 2015. doi:10.1088/1748-0221/10/04/C04039.

[8] The CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," *Phys. Lett. B*, vol. 716, 2012. doi:10.1016/j.physletb.2012.08.021.

[9] The CMS and LHCb Collaborations, "Observation of the rare $B_s^0 \to \mu^+\mu^-$ decay from the combined analysis of CMS and LHCb data," *Nature*, vol. 522, 2015. doi:10.1038/nature14474.

[10] R. Calaga, "Crab cavities for the LHC upgrade," *CERN Document Server*, 2012. doi:10.5170/CERN-2012-006.

[11] P. Campana, M. Klute, and P.S. Wells, "Physics Goals and Experimental Challenges of the Proton-Proton High-Luminosity Operation of the LHC," *Annual Review of Nuclear and Particle Science*, 2016. doi:10.1146/annurev-nucl-102115-044812.

[12] M. Fabbrichesi et al., "Vector boson scattering at the LHC. A study of the WW -¿ WW channels with the Warsaw cut," *Physical Review Letters*, 2015. doi:10.1103/PhysRevD.93.015004.

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

[14] CMS collaboration, "Technical Proposal for the Phase-II upgrade of the Compact Muon Solenoid," tech. rep., CERN, 2015.

[15] J.-B. Sauvan, "Concepts and design of the CMS High Granularity Calorimeter Level 1 Trigger," *CALOR2016 17th International Conference on Calorimetry in Particle Physics*, 2016.

[16] M. Di Nardo, "The pixel detector for the CMS phase-II upgrade," *Journal of Instrumentation*, 2014. doi:10.1088/1748-0221/10/04/C04019.

[17] G. Steinbrück, "Small pitch pixel sensors for the CMS phase II upgrade," *IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, 2015. doi:10.1109/NSSMIC.2015.7581861.

[18] http://www.xray-imatek.com/technology/hybryd-pixel-detectors.

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

[20] H. Spieler, "Lectures on Detector Techniques," 1999.

[21] K. Motohashi, "Evaluation of KEK n-in-p planar pixel sensor structures for very high radiation environments with testbeam," *Nuclear Instruments and Methods A*, 2014. doi:10.1016/j.nima.2014.05.092.

[22] G.F. Dalla Betta et al., "Development of a new generation of 3D pixel sensors for HL-LHC," *Nuclear Instruments and Methods A*, 2015. doi:10.1016/j.nima.2015.08.032.

[23] C. Da Via et al., "3D silicon sensors: Design, large area production and quality assurance for the ATLAS IBL pixel detector upgrade," *Nuclear Instruments and Methods A*, 2012. doi:10.1016/j.nima.2012.07.058.

[24] J. Christiansen and M. Garcia Sciveres, "2013 RD Collaboration proposal: Development of pixel readout integrated circuits for extreme rate and radiation," tech. rep., CERN, 2013.

[25] N. Demaria et al., "Recent progress of RD53 Collaboration towards next generation Pixel Read-Out Chip for HL-LHC," *Journal of Instrumentation*, 2016. doi:10.1088/1748-0221/11/12/C12058.

[26] Y. Tsividis, *Operation and Modeling of the MOS Transistor*. McGraw Hill, 1999.

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

[28] P. E. Allen and D. R. Holberg, *CMOS Analog Circuit Design - Second Edition*. Oxford University Press, 2002.

[29] A. Rivetti, *CMOS: Front-End Electronics for Radiation Sensors*. CRC Press, 2015.

[30] A. J. M. Rabaey and B. Nikolic, *Digital Integrated Circuits: A Design Perspective - Second Edition*. Prentice Hall, 2002.

[31] B. Razavi, *Fundamentals of Microelectronics - Preview Edition*. Wiley, 2006.

[32] F. Maloberti, *Analog Design for CMOS VLSI Systems*. Kluwer Academic Publisher, 2003.

[33] C. C. Enz and E. A. Vittoz, *Charge-Based MOS Transistor Modeling: The EKV Model for Low-Power and RF IC Design*. Wiley, 2006.

[34] F. Silveira, D.Flandre, and P. G. A. Jespers, "A g_m/I_D Based Methodology for the Design of CMOS Analog Circuits and Its Application to the synthesis of a Silicon-on-Insulator Micropower OTA," *IEEE Journal of solid-state physics*, 1996. doi:10.1109/4.535416.

[35] S. Kang and Y. Leblebici, *CMOS Digital Integrated Circuits: Analysis and Design*. McGraw Hill, 2003.

[36] F. Resta et al., "IC-PIX28: A 28nm read-out channel for pixel detector," *Electronics, Circuits, and Systems (ICECS), 2015 IEEE International Conference on Electronics, Circuits, and Systems (ICECS)*, 2015. doi:10.1109/ICECS.2015.7440329.

[37] K.R. Lakshmikumar et al., "Characterisation and modeling of mismatch in MOS transistors for precision analog design," *IEEE Journal of Solid-State Circuits*, 1986. doi:10.1109/JSSC.1986.1052648.

[38] M.J.M. Pelgrom et al., "Matching properties of MOS transistors," *IEEE Journal of Solid-State Circuits*, 1986. doi:10.1109/JSSC.1989.572629.

[39] M.J.M. Pelgrom et al., "Transistor matching in analog CMOS applications," *Electron Devices Meeting, 1998. IEDM '98. Technical Digest., International*, 1998. doi:10.1109/IEDM.1998.746503.

[40] I. Sanchez Esqueda, *Modeling of Total Ionizing Dose Effects in Advanced Complementary Metal-Oxide-Semiconductor Technologies.* PhD thesis, Arizona University, 2011.

[41] F. Faccio. $https://lhcb-elec.web.cern.ch/lhcb-elec/papers/radiation_tutorial.pdf$, 2005.

[42] N. S. Saks et al., "Generation of interface states by ionizing radiation in very thin MOS oxides," *IEEE Transactions on Nuclear Science*, 1986. doi:10.1109/TNS.1986.4334576.

[43] F. Faccio and G. Cervelli, "Radiation-induced edge effects in deep submicron CMOS transistors," *IEEE Transactions on Nuclear Science*, 2005. doi:10.1109/TNS.2005.860698.

[44] Fan Xue et al., "Gate-enclosed NMOS transistors," *Journal of Semiconductor*, 2011. doi:10.1088/1674-4926/32/8/084002.

[45] W. Snoeys et al., "Layout techniques to enhance the radiation tolerance of standard CMOS technologies demonstrated on a pixel detector readout chip," *Nuclear Instruments and Methods A*, 1999. doi:10.1016/S0168-9002(99)00899-2.

[46] M. Menouni et al., "1-Grad total dose evaluation of 65 nm CMOS technology for the HL-LHC upgrades," *Journal of Instrumentation*, 2015. doi:10.1088/1748-0221/10/05/C05009.

[47] H. Spieler, "Analog and Digital Electronics for Detectors," *Proceedings of the 2003 ICFA School on Instrumentation*, 2003.

[48] E. Simoen, "On the flicker noise in submicron silicon MOSFETs," *Solid-State Electronics*, 1999. doi:10.1016/S0038-1101(98)00322-0.

[49] L. Cristofolini. $http://www.fis.unipr.it/\ gigi/dida/strumentazione/harvard_noise.pdf$, 2004.

[50] D. Bortoletto, "How and why silicon sensors are becoming more and more intelligent?," *Journal of Instrumentation*, 2015. doi:10.1088/1748-0221/10/08/C08016.

[51] C. Adolphsen, "The Mark II Vertex Detector, a perfect detector for an imperfect collider," 2012.

[52] A. Litke et al., "A silicon strip vertex detector for the Mark II experiment at the SLAC linear collider," *Nuclear Instruments and Methods in Physics Research Section A*, 1988. doi:10.1016/0168-9002(88)91059-5.

[53] G. Anzivino et al., "First results from a silicon-strip detector with VLSI readout," *Nuclear Instruments and Methods in Physics Research Section A*, 1986. doi:10.1016/0168-9002(86)90835-1.

[54] J. Walker et al., "Development of high density readout for silicon strip detectors," *Nuclear Instruments and Methods in Physics Research Section A*, 1984. doi:10.1016/0168-9002(84)90191-8.

[55] G. Anzivino et al., "Latest results from silicon microstrip detectors with VLSI readout for the delphi microvertex detector," *Nuclear Instruments and Methods in Physics Research Section A*, 1987. doi:10.1016/0168-9002(87)91039-4.

[56] A. Breakstone et al., "Radiation Hardness and Annealing Tests of a Custom VLSI Device," *IEEE Transactions on Nuclear Science*, 1987. doi:10.1109/TNS.1987.4337391.

[57] P. Seller et al., "Results of silicon strip detector readout using a CMOS low power microplex (MX1)," *IEEE Transactions on Nuclear Science*, 1988. doi:10.1109/23.12700.

[58] E. Beuville et al., "Amplex, a low-noise, low-power analog CMOS signal processor for multi-element silicon particle detectors," *Nuclear Instruments and Methods in Physics Research Section A*, 1990. doi:10.1016/0168-9002(90)90481-K.

[59] R. Ansari et al., "The silicon detectors in the UA2 experiment," *Nuclear Instruments and Methods in Physics Research Section A*, 1989. doi:10.1016/0168-9002(89)91111-X.

[60] W. Snoeys et al., "Pixel readout chips in deep submicron CMOS for ALICE and LHCb tolerant to 10Mrad and beyond," *Nuclear Instruments and Methods in Physics Research Section A*, 2001. doi:10.1016/S0168-9002(01)00590-3.

[61] R. Dinapoli et al., "A front-end for silicon pixel detectors in ALICE and LHCb," *Nuclear Instruments and Methods in Physics Research Section A*, 2000. doi:10.1016/S0168-9002(00)01281-X.

[62] R. Szczygiel et al., "A Prototype Pixel Readout IC for High Count Rate X-Ray Imaging Systems in 90 nm CMOS Technology," *IEEE Transactions on Nuclear Science*, 2010. doi:10.1109/TNS.2010.2044664.

[63] I. Peric et al., "The FEI3 readout chip for the ATLAS pixel detector," *Nuclear Instruments and Methods in Physics Research Section A*, 2006. doi:10.1016/j.nima.2006.05.032.

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

[65] L. Pacher, *Development of Integrated Pixel Front-End Electronics in 65 nm CMOS Technology for Extreme Rate and Radiation at HL-LHC*. PhD thesis, Università di Torino, 2015.

[66] T. Kugathasan, *Low-Power High Dynamic Range Front-End Electronics for the Hybrid Pixel Detectors of the PANDA MVD*. PhD thesis, Università di Torino, 2011.

[67] *Abort Gap Cleaning for LHC Run 2*, 2014.

[68] Y. Huang, H. Schleifer, and D. Killat, "Design and analysis of novel dynamic latched comparator with reduced kickback noise for high-speed ADCs," *IEEE ECCTD*, 2015. doi:10.1109/ECCTD.2013.6662236.

[69] A. Kruth et al., "GOSSIPO-3: measurements on the prototype of a read-out pixel chip for Micro-Pattern Gaseous Detectors," *Journal of Instrumentation*, 2010. doi:10.1088/1748-0221/5/12/C12005.

[70] C.-C. Liu, S.-J. Chang, G.-Y. Huang, and Y.-Z. Lin, "A 10-bit 50-MS/s SAR ADC With a Monotonic Capacitor Switching Procedure," *IEEE Journal of solid-state circuits*, vol. 45, 2010. doi:10.1109/JSSC.2010.2042254.

[71] M. Badaroglu et al., "Evolution of Substrate Noise Generation Mechanisms With CMOS Technology Scaling," *IEEE Transaction on circuits and systems I*, 2006. doi:10.1109/TCSI.2005.856049.

[72] E. Monteil, N. Demaria, L. Pacher, A. Rivetti, M. Da Rocha Rolo, F. Rotondo, and C. Leng, "Pixel front-end with synchronous discriminator and fast charge measurement for the upgrades of HL-LHC experiments," *Journal of Instrumentation*, 2016. doi:10.1088/1748-0221/11/03/C03013.

[73] F. Faccio et al., "Radiation-Induced Short Channel (RISCE) and Narrow Channel (RINCE) Effects in 65 and 130 nm MOSFETs," *IEEE Transaction of Nuclear Science*, 2015. doi:10.1109/TNS.2015.2492778.

[74] L. M. Jara Casas et al., "Characterization of radiation effects in 65 nm digital circuits with the DRAD digital radiation test chip," *Journal of Instrumentation*, vol. 12, 2017. doi:10.1088/1748-0221/12/02/C02039.

[75] E. Monteil et al., "A Prototype of a New Generation Readout ASIC in 65nm CMOS for Pixel Detectors at HL-LHC," *Journal of Instrumentation*, 2016. doi:10.1088/1748-0221/11/12/C12044.

[76] L. Gaioni et al., "65 nm CMOS analog front-end for pixel detectors at the HL-LHC," *Journal of Instrumentation*, 2016. doi:10.1088/1748-0221/11/02/C02049.

[77] G. De Robertis, F. Loddo, S. Mattiazzo, L. Pacher, D. Pantano, and C. Tamma, "Design of a 10-bit segmented current-steering digital-to-analog converter in CMOS 65 nm technology for the bias of new generation readout chips in high radiation environment," *Journal of Instrumentation*, 2016. doi:10.1088/1748-0221/11/01/C01027.

[78] G. Traversi et al., "Characterization of bandgap reference circuits designed for high energy physics applications," *Nuclear Instruments and Methods in Physics Research Section A*, 2016. doi:10.1016/j.nima.2015.09.103.

[79] M. Garcia Sciveres et al., "The FE-I4 pixel readout integrated circuit," *Nuclear Instruments and Methods in Physics Research Section A*, 2011. doi:10.1016/j.nima.2010.04.101.

[80] M. Garcia Sciveres et al., "Results of FE65-P2 Pixel Readout Test Chip for High Luminosity LHC Upgrades," *Proceedings of Science*, 2016.